SAMSI Summer 2015: CCNS
Computational Neuroscience Summer School

SPIKE TRAIN ANALYSIS USING GENERALIZED LINEAR MODELS

Uri Eden

BU Department of Mathematics and Statistics

July 27, 2015

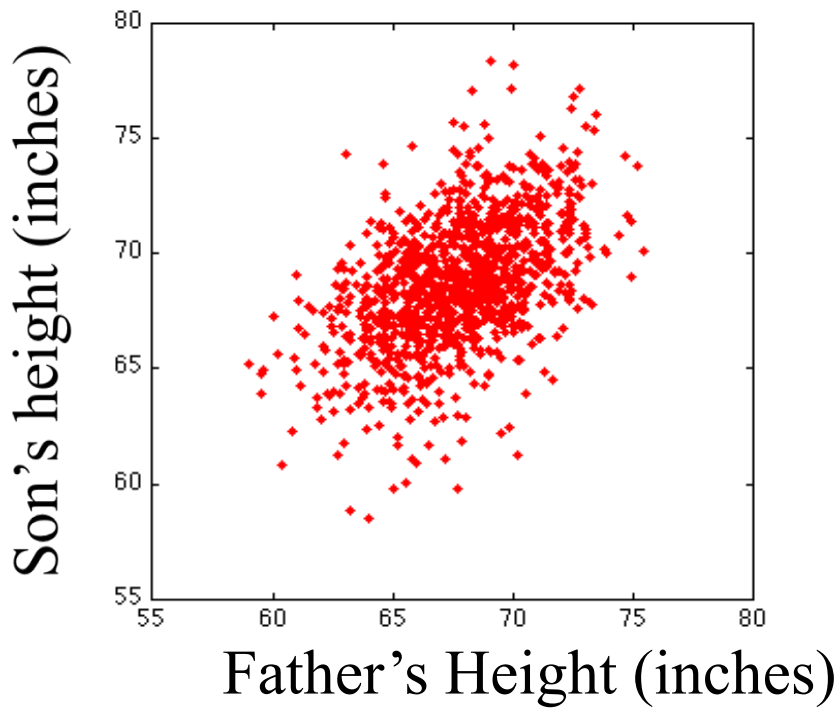
Outline

- Quick introduction to GLM theory
- GLM model for inhomogeneous Poisson spiking
- History dependent GLM model of retinal neurons in culture
- A GLM model of learning in behavioral experiments

Simple linear regression

- How does the height of a son depend of the height of his father?

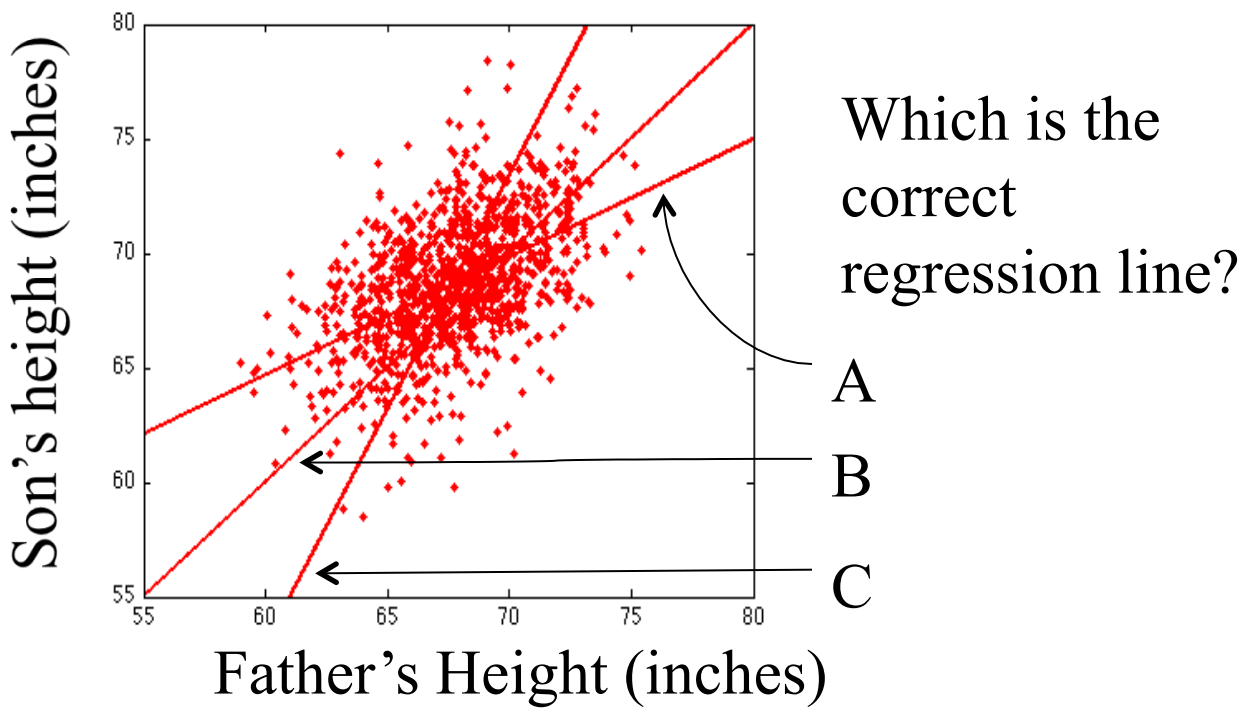
$$S_i = bF_i + \varepsilon_i$$



Simple linear regression

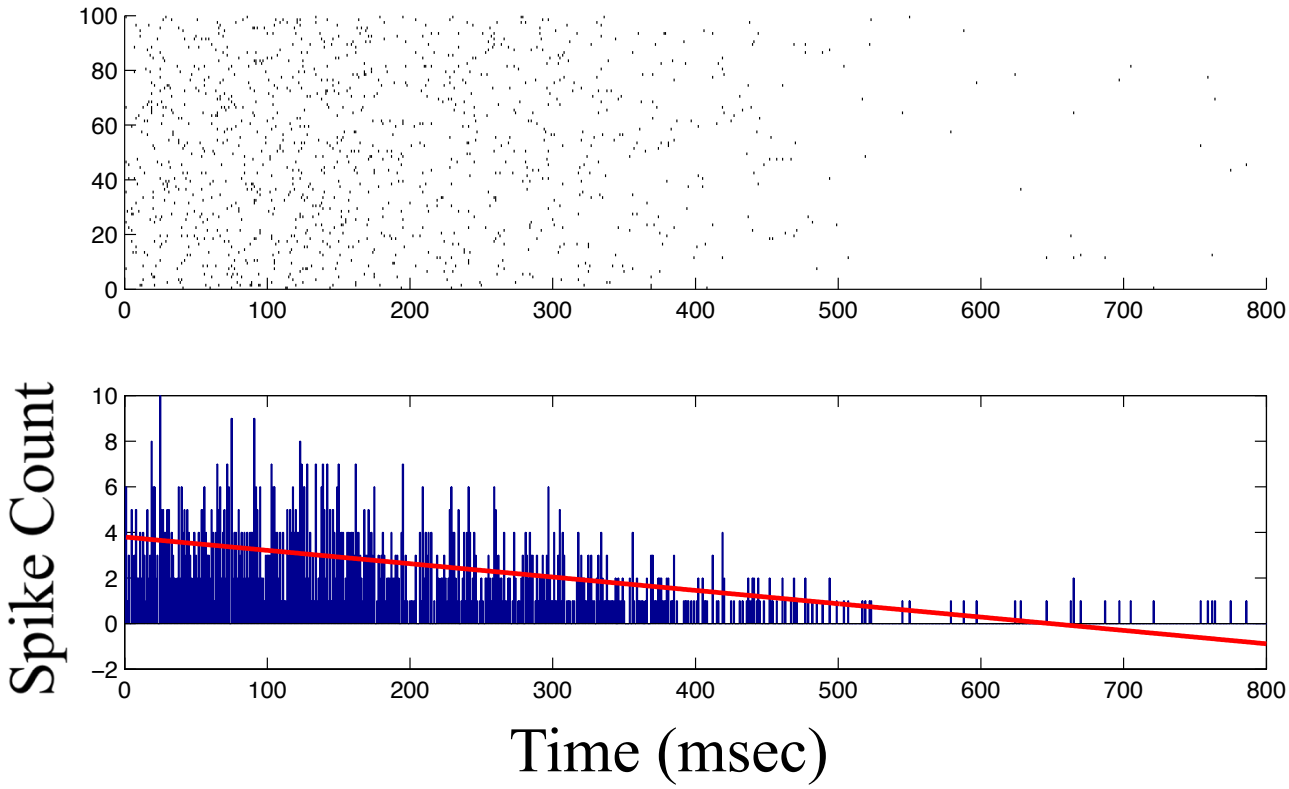
- How does the height of a son depend of the height of his father?

$$S_i = bF_i + \varepsilon_i$$



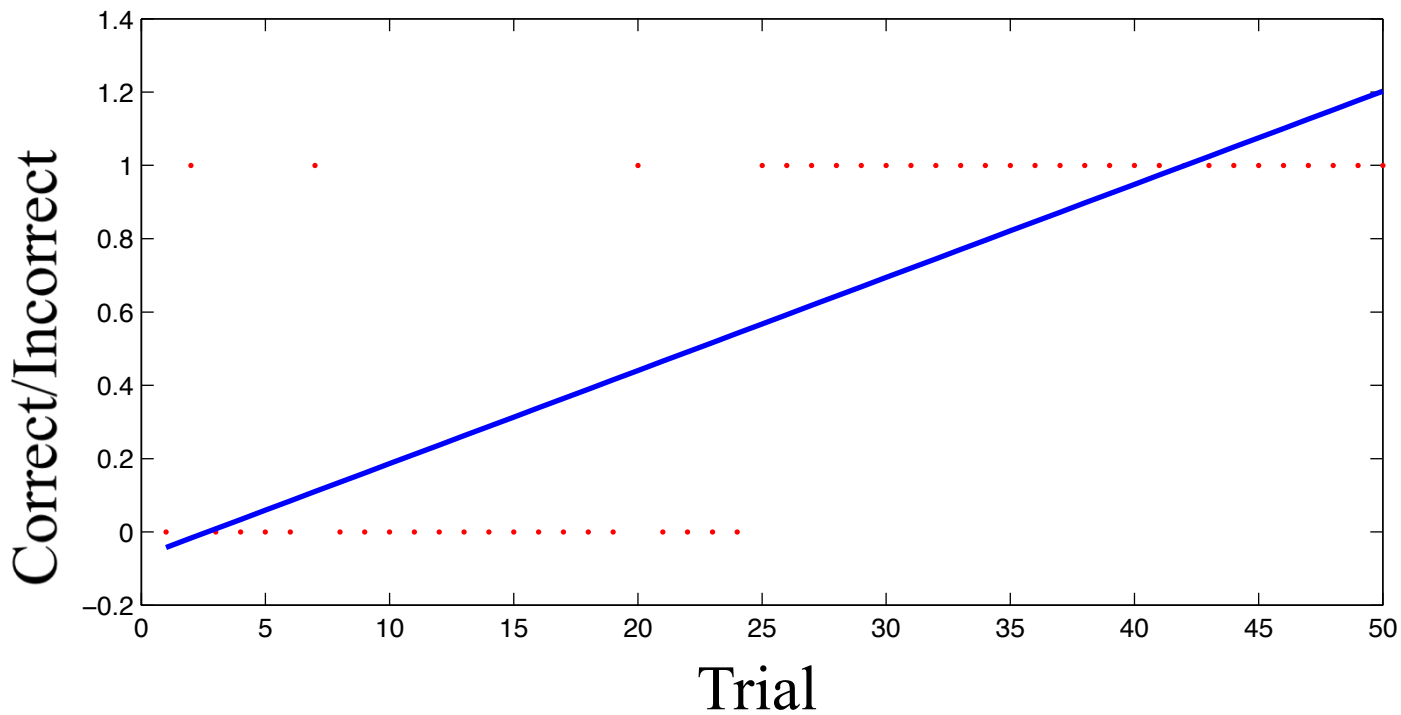
Count data

- Linear regression methods are not well suited for count data



Binary data

- Linear regression methods are not well suited for binary data



Generalized Linear Models

- Linear regression models of the form:

$$y = X\beta + \varepsilon \qquad \varepsilon \sim N(0, \Sigma)$$

  are useful for relating continuous valued observations to a set of covariates.
- Many types of data cannot be described by a Gaussian additive noise model.
- Generalized linear models extend a simple class of models to many additional data types.

Count data:

$$\log\left(\hat{\lambda}\right) = X\beta$$

Binary data:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = X\beta$$

The Natural Exponential Family

A probability model for the data $\{y_1, \ldots, y_n\}$ is in the exponential family if you can write its likelihood in the form:

$$L = \prod_{Data} \exp\{T(y_i)C(\theta) + H(y_i) + D(\theta)\}$$

Some common distributions in the exponential family include:

- Normal, Bernoulli, binomial, Poisson, gamma, beta, exponential, chi-square, lognormal, …

Generalized Linear Models

$$L = \prod_{Data} \exp\{T(y_i)C(\theta) + H(y_i) + D(\theta)\}$$

Set the link function, $C(\theta)$, to be a linear function of the covariates,

$$C(\theta) = \alpha_0 + \sum_{j=1}^{p} \alpha_j g_j\left(X_i\right)$$

Differentiate the log likelihood with respect to the parameters, set equal to zero, and solve the resulting system of equations of the form

$$\sum T(y_i) g_j(X_i) + \left.\frac{\partial D}{\partial \alpha_j}\right|_{\hat{\theta}} = 0$$

The Natural Exponential Family

$$L = \prod_{Data} \exp\{T(y_i)C(\theta) + H(y_i) + D(\theta)\}$$

Poisson Data:

$$L = \prod_{Data} \frac{\lambda_k^{\,y_k} \exp\{-\lambda_k\}}{y_k\,!}$$

$$= \prod_{Data} \exp\{y_k \log(\lambda_k) - \log(y_k\,!) - \lambda_k\}$$

So the link function is:

$$C(\theta) = \log\left(\lambda_k\right)$$

The Natural Exponential Family

$$L = \prod_{Data} \exp\{T(y_i)C(\theta) + H(y_i) + D(\theta)\}$$

Binomial Data:

$$L = \prod_{Data} p_k^{\,y_k} (1 - p_k)^{1-y_k}$$

$$= \prod_{Data} \exp\{y_k \log(p_k) + (1 - y_k)\log(1 - p_k)\}$$

$$= \prod_{Data} \exp\{y_k \log(\frac{p_k}{1 - p_k}) + \log(1 - p_k)\}$$

So the link function is:

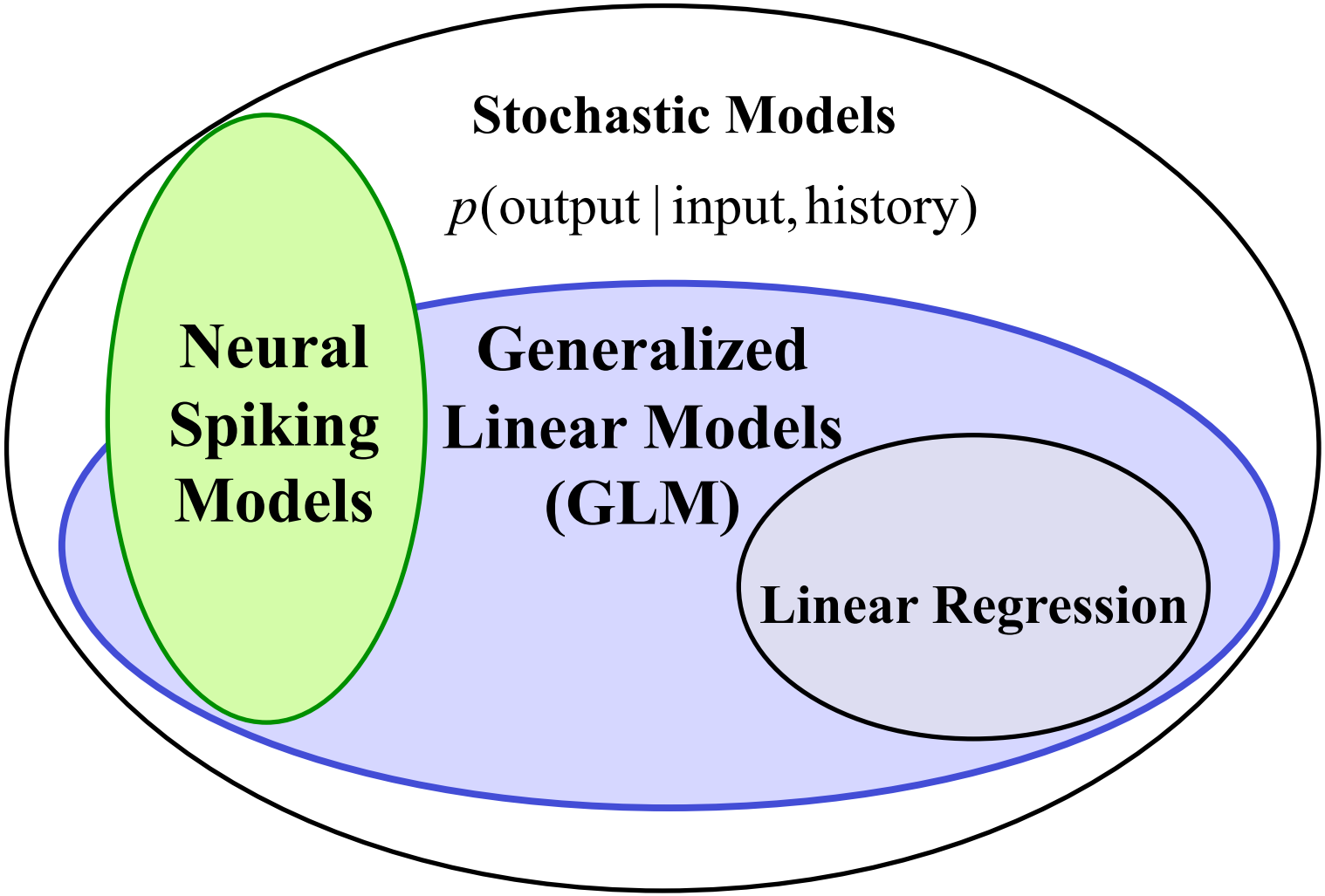
$$C(\theta) = \log\left(\frac{p_k}{1 - p_k}\right)$$

GLM Models for Spike Data

Link Function	Distribution	Equation
logit	Binomial	$\log\left(\frac{p_k}{1-p_k}\right) = \alpha_0 + \sum_{j=1}^{p} \alpha_j g_j\left(X_k\right)$
log	Poisson	$\log\left(\lambda_k\right) = \alpha_0 + \sum_{j=1}^{p} \alpha_j g_j\left(X_k\right)$

Fitting GLM

- As with ISI models, use maximum likelihood to obtain GLM parameters.
- In general it is not possible to obtain a closed form solution for the ML estimator or for its distribution.
- So, use your favorite numerical optimization technique (such as Newton's method), or my favorite: MATLAB™

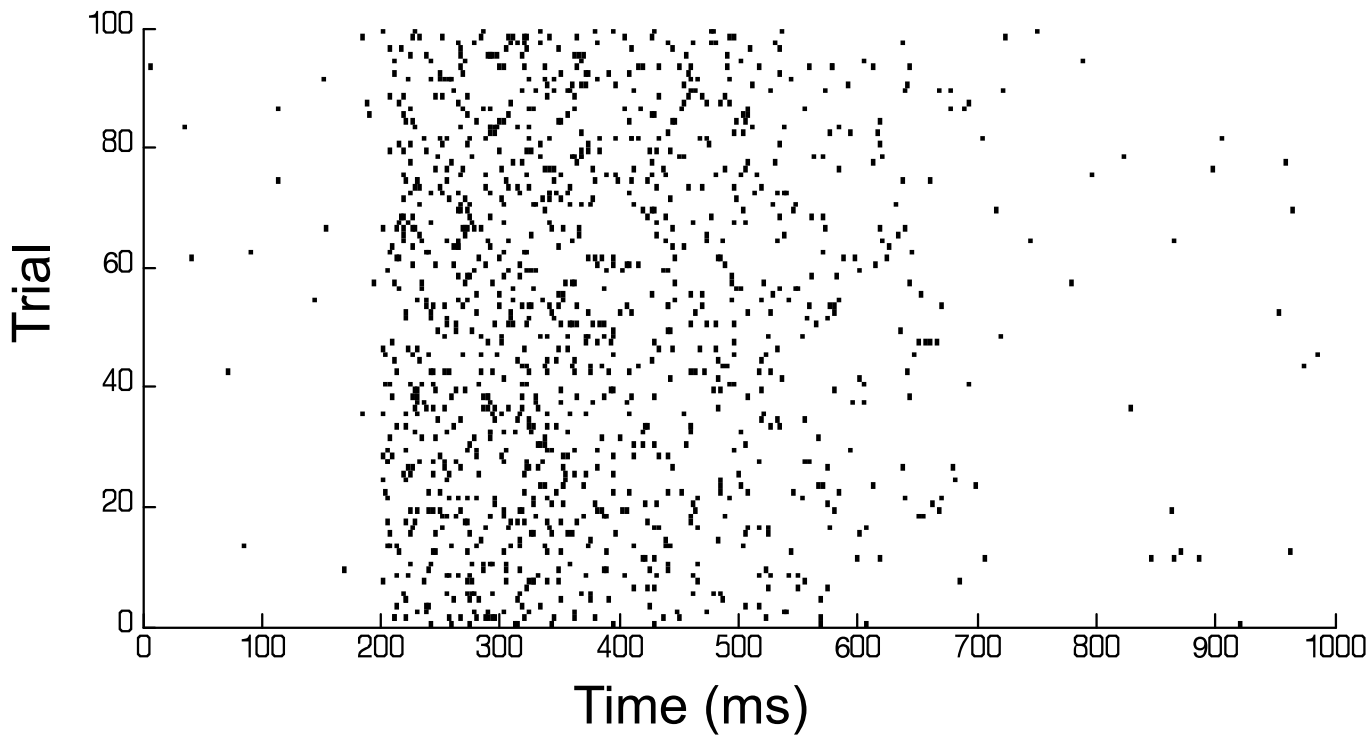


Properties of GLM:

- Convex likelihood surface
- Estimators asymptotically have minimum MSE

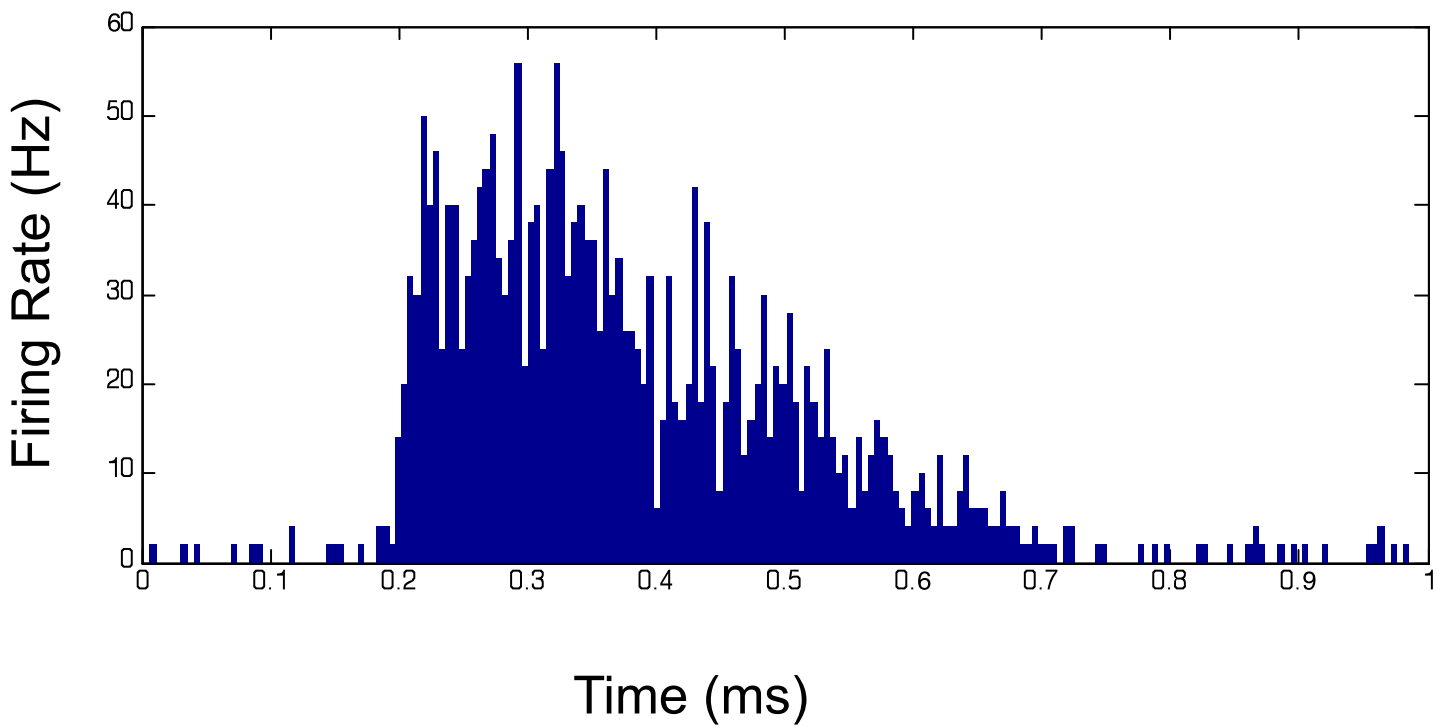
Case 1: Inhomogeneous Poisson Model

- Construct an inhomogeneous Poisson spiking model for repeated trial data as a function of time



Case 1: Inhomogeneous Poisson Model

- Construct an inhomogeneous Poisson spiking model for repeated trial data as a function of time



Case 1: Inhomogeneous Poisson Model

- For an inhomogeneous Poisson model for repeated trial data as a function of time
  - Polynomial model:

$$\log(\lambda_t) = \alpha_0 + \sum_{j=1}^{p} \alpha_j t^j$$

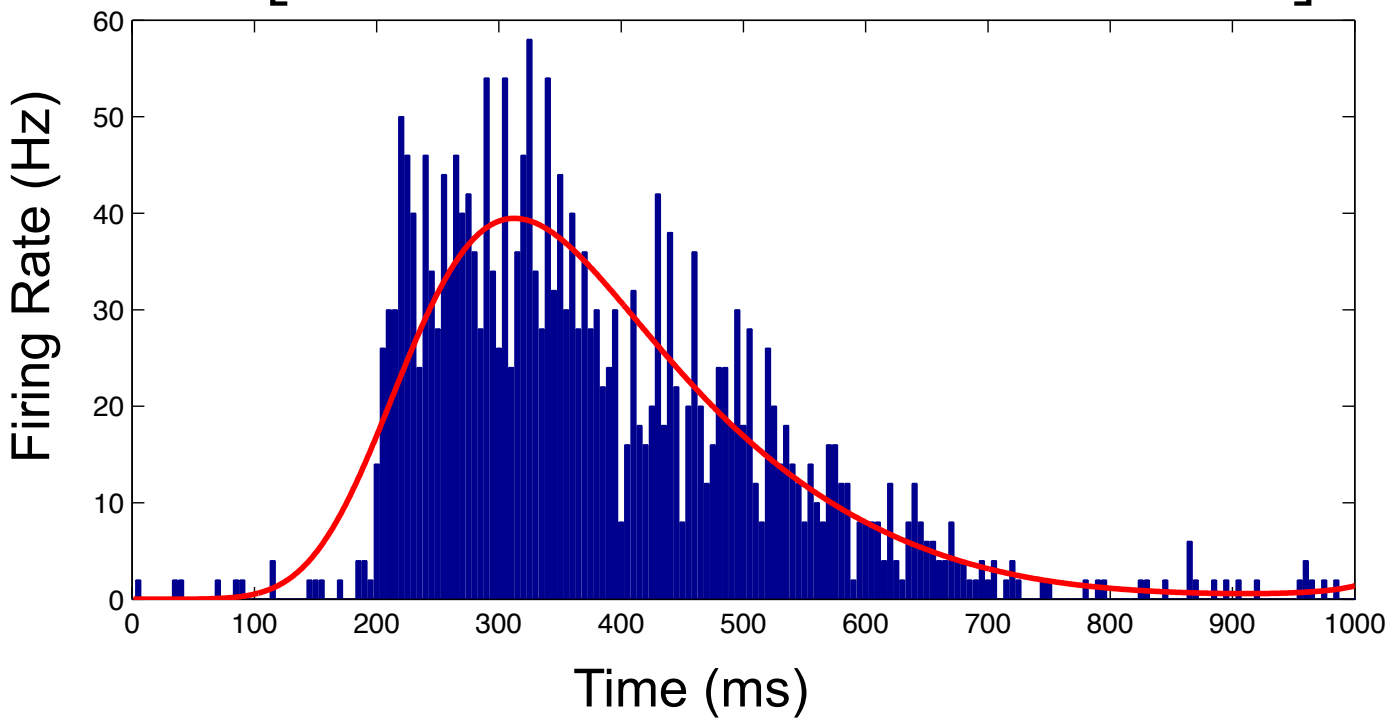
or:

$$\lambda_t = e^{\alpha_0 + \sum_{j=1}^{p} \alpha_j t^j}$$

Case 1: Inhomogeneous Poisson Model

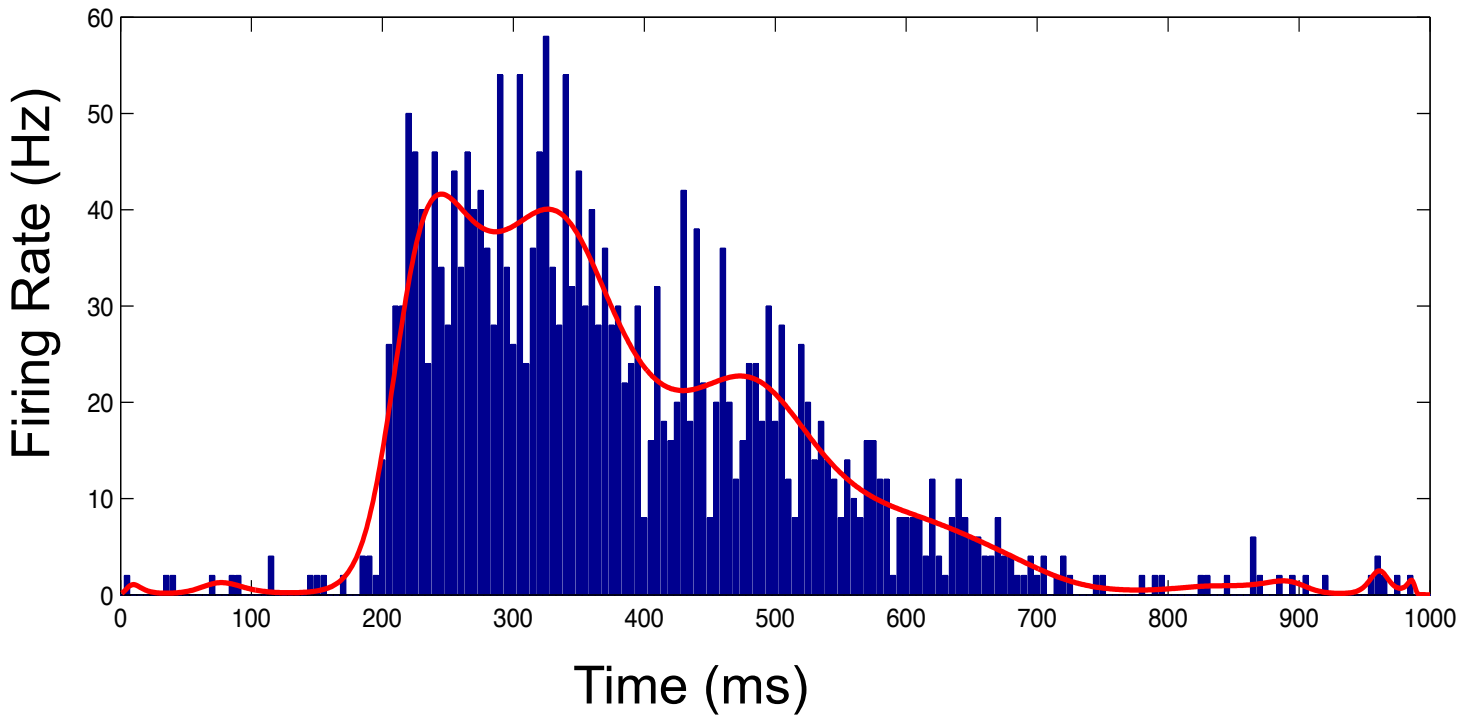
- Inhomogeneous Poisson GLM using 5th order polynomial in time

$$\hat{\alpha} = \begin{bmatrix} -11 & 126 & -462 & 815 & -722 & 252 \end{bmatrix}$$



Case 1: Inhomogeneous Poisson Model

- Inhomogeneous Poisson GLM using 50th order polynomial in time



Goodness-of-fit measures

Akaike's Information Criterion:

$$-2\log f(w \mid \hat{\theta}_{ML}) + 2p$$

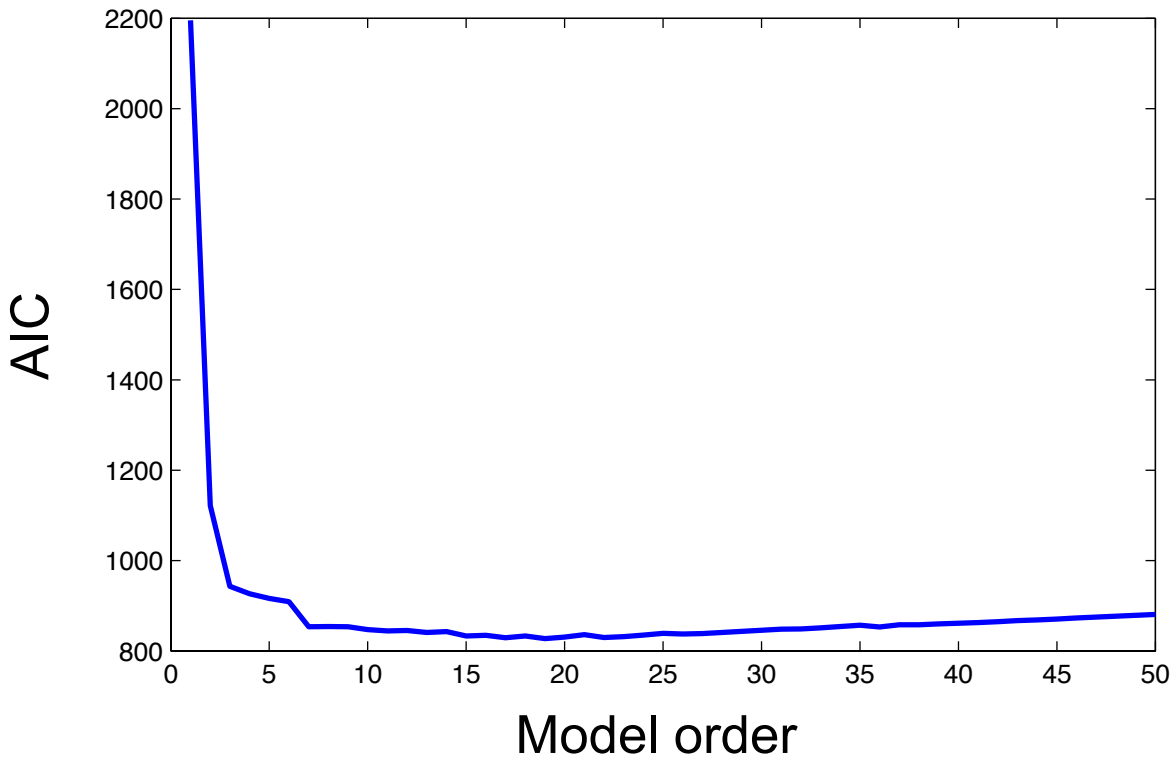
For maximum likelihood estimates it measures the trade-off between maximizing the likelihood (minimizing $-2\log f(w \mid \hat{\theta}_{ML})$) and the numbers of parameters p, the model requires.

Selecting the (parsimonious) model that minimizes the AIC:

- **Helps prevent overfitting**
- **Is asymptotically equivalent to complete leave-one-out cross-validation**
- **Asymptotically minimizes the KL distance between the selected model and the true unknown model**

Case 1: Inhomogeneous Poisson Model

- AIC plot of model order



Case 1: Inhomogeneous Poisson Model

- For an inhomogeneous Poisson model for repeated trial data as a function of time
  - Spline model:

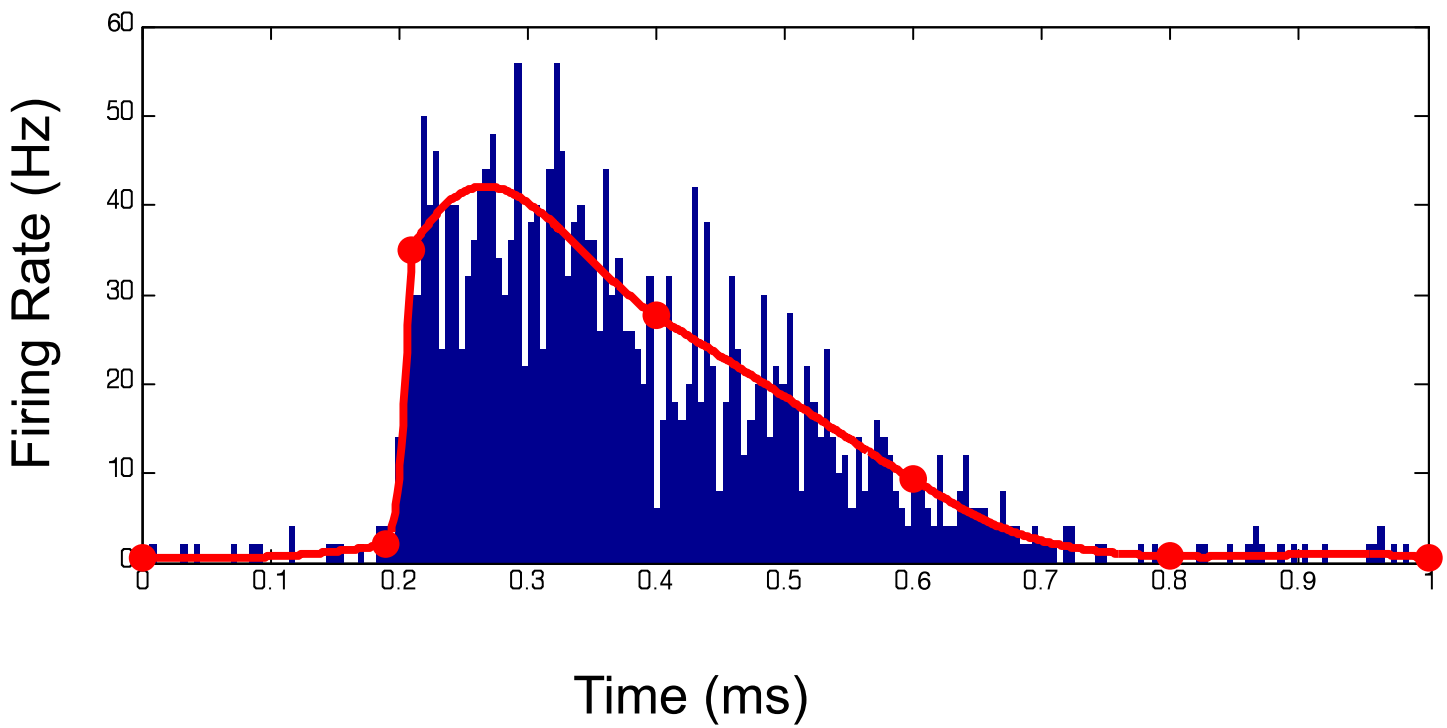
$$\lambda_t = e^{\sum_{j=1}^{p} \alpha_j c_j(t)}$$

where $c_j(t)$ are spline basis functions

Case 1: Inhomogeneous Poisson Model

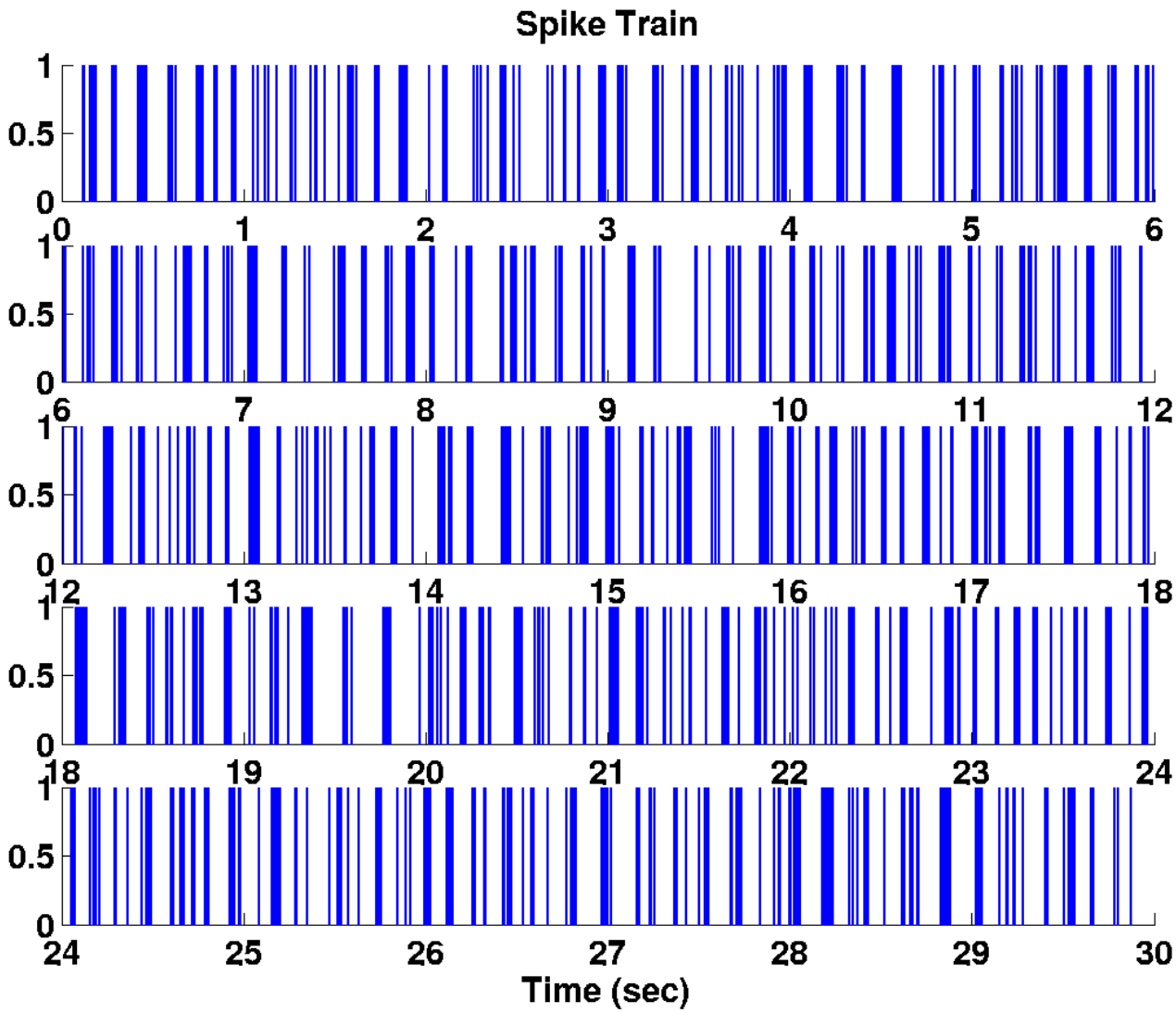
- Inhomogeneous Poisson GLM using spline fit in time

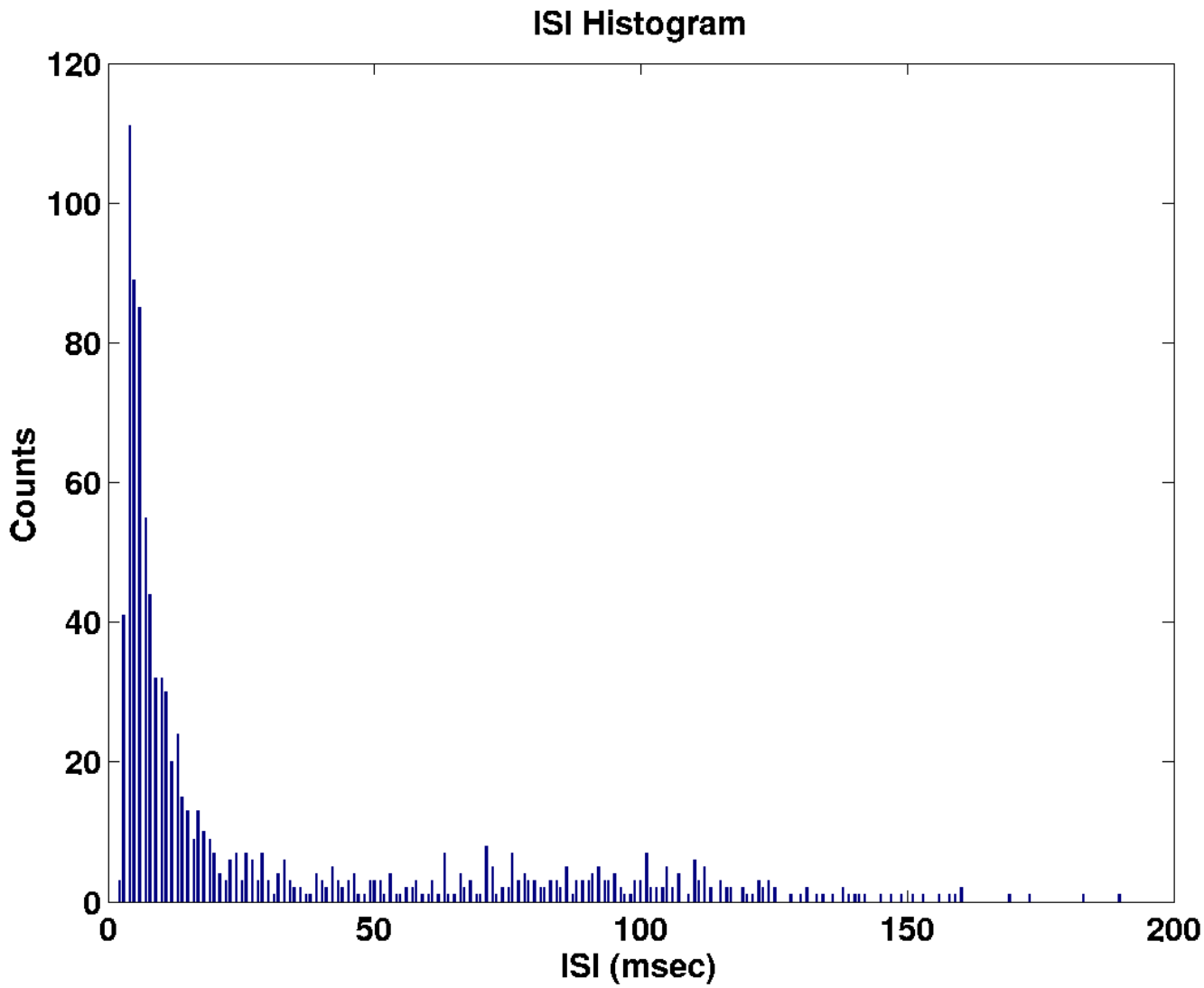
$$\exp(\hat{\alpha}) = \begin{bmatrix} .5 & 1 & 30 & 27 & 9 & 1 & .5 \end{bmatrix}$$

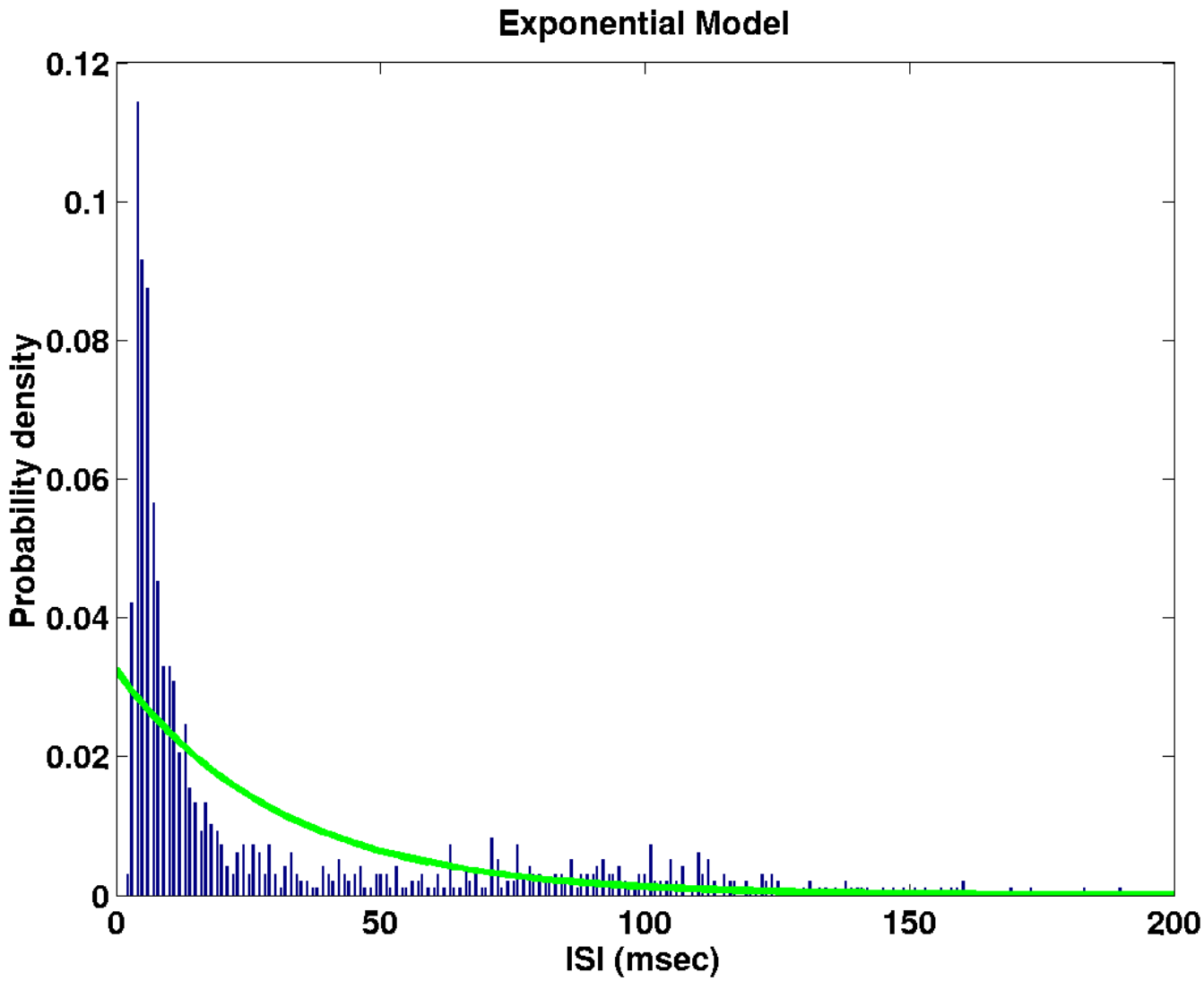


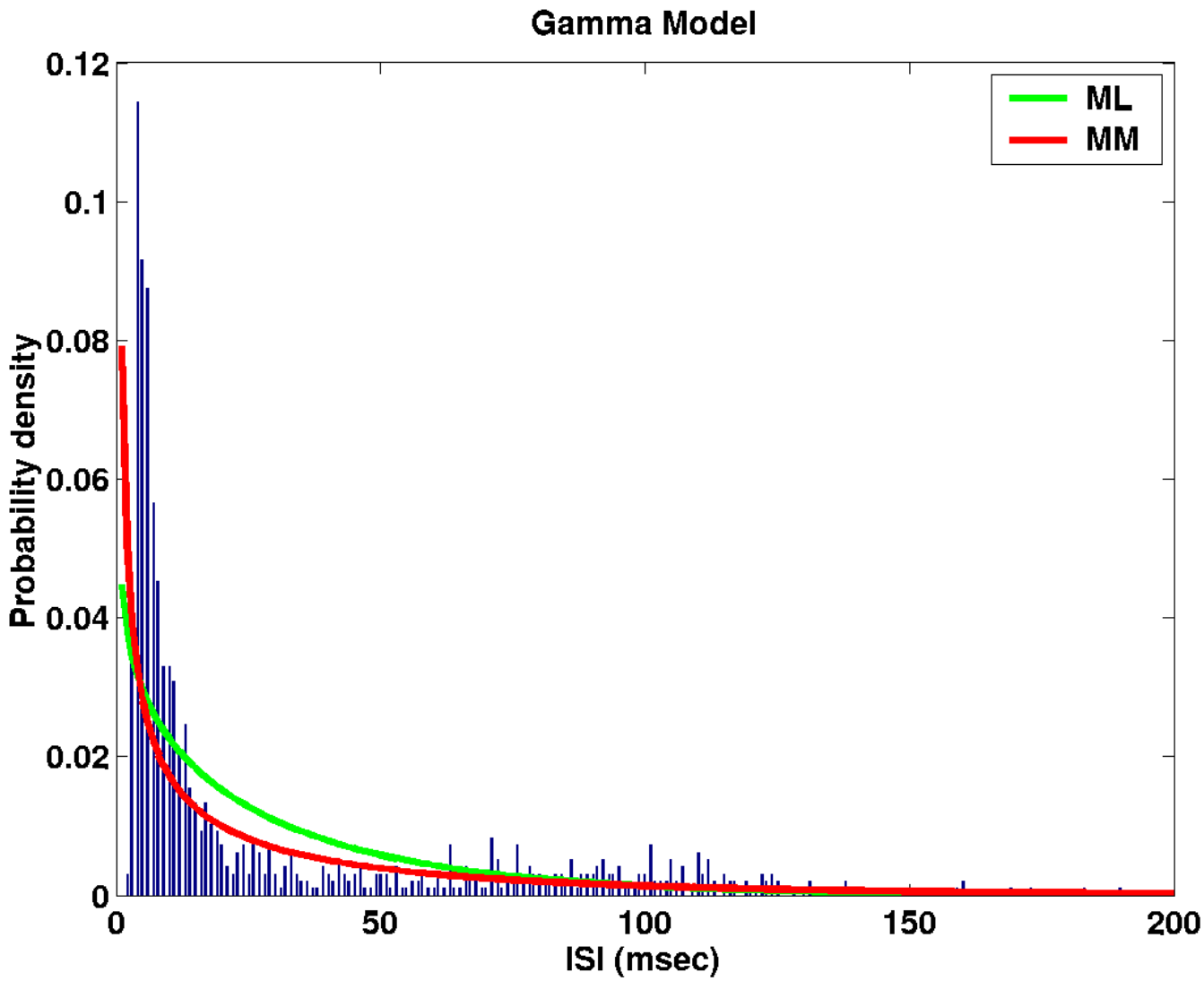
Case 2: An Analysis of the Spiking Activity of Retinal Neurons in Culture (Iygengar and Liu, 1997)

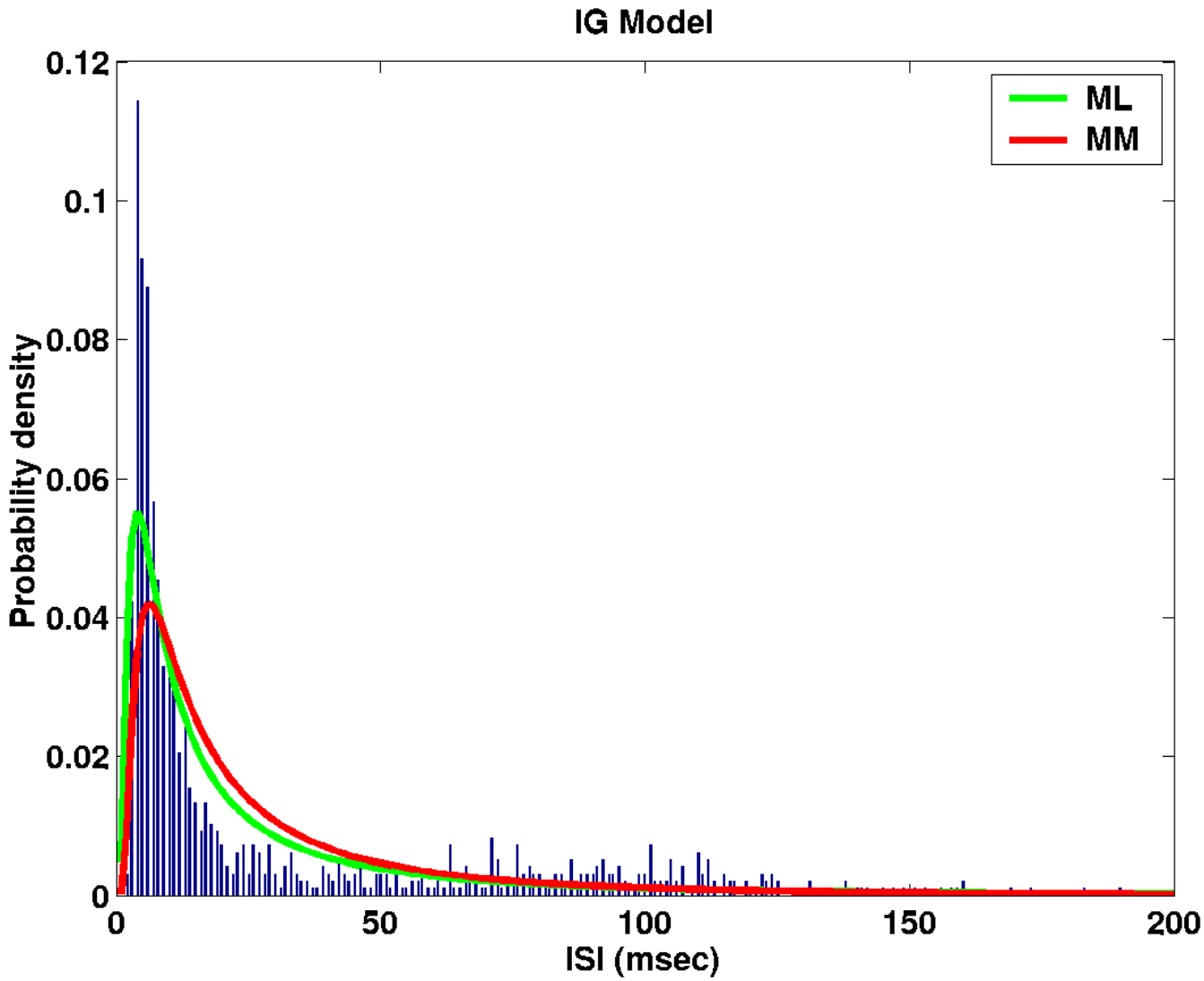
Retinal neurons are grown in culture under constant light and environmental conditions. The spontaneous spiking activity of these neurons is recorded. The objective is to develop a statistical model which accurately describes the stochastic structure of this activity.


Spike Train
1
0.5
0
0
1
2
3
4
5
6
6
7
8
9
10
11
12
12
13
14
15
16
17
18
18
19
20
21
22
23
24
24
25
26
27
28
29
30
Time (sec)

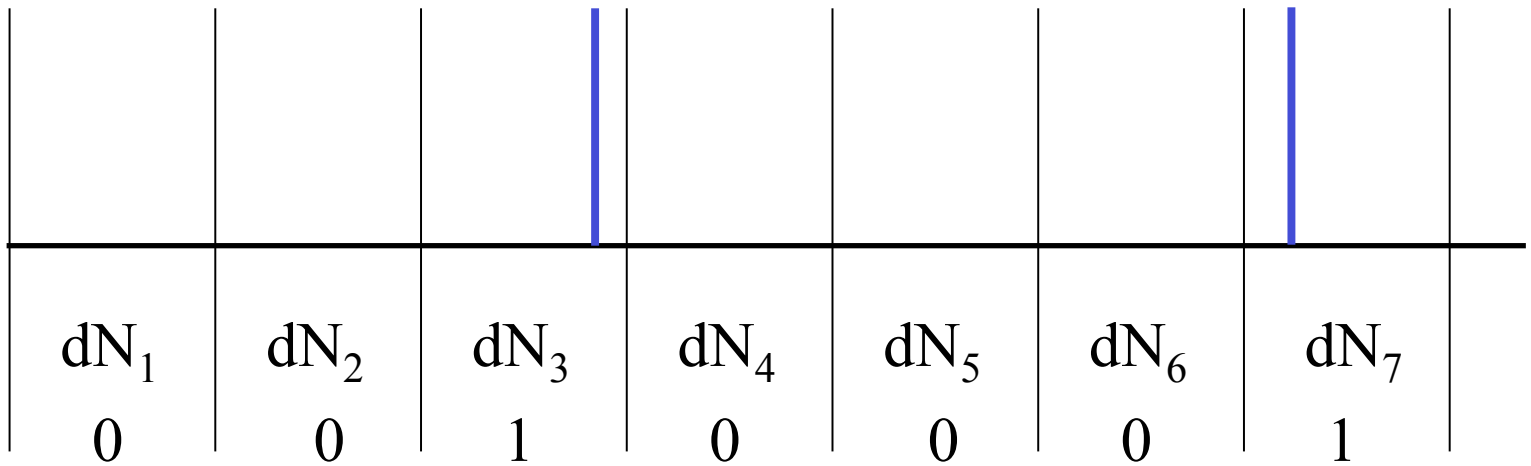

ISI Histogram
Counts
ISI (msec)
0
20
40
60
80
100
120
0
50
100
150
200


Exponential Model
Probability density
0.12
0.1
0.08
0.06
0.04
0.02
0
0
50
100
150
200
ISI (msec)


Gamma Model
ML
MM
Probability density
ISI (msec)
0.12
0.1
0.08
0.06
0.04
0.02
0
0
50
100
150
200


IG Model
ML
MM
Probability density
ISI (msec)
0.12
0.1
0.08
0.06
0.04
0.02
0
0
50
100
150
200

Discrete Time Spike Train Data



dN_k is the spike indicator function in interval k

λ_k is the intensity of spiking at time k, which in the limit is given by

$$\lambda(t \mid H_t) = \lim_{\Delta t \to 0} \frac{\Pr(\text{Spike in } (t, t+\Delta t) \mid H_t)}{\Delta t}$$

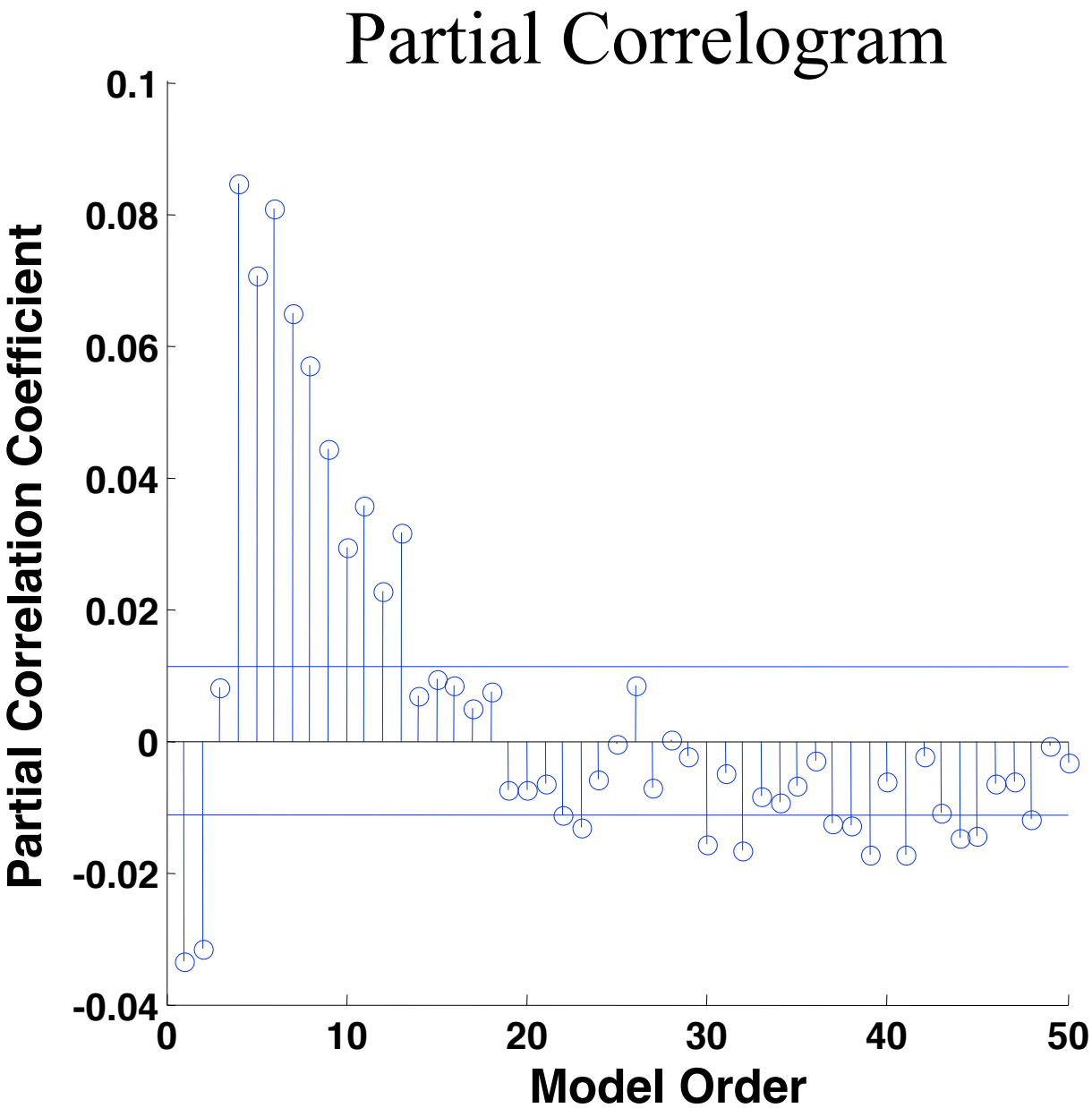
GLM History Model

The ISI distribution models we constructed assume that $p(\mathrm{ISI} \mid H_t) = p(\mathrm{ISI})$

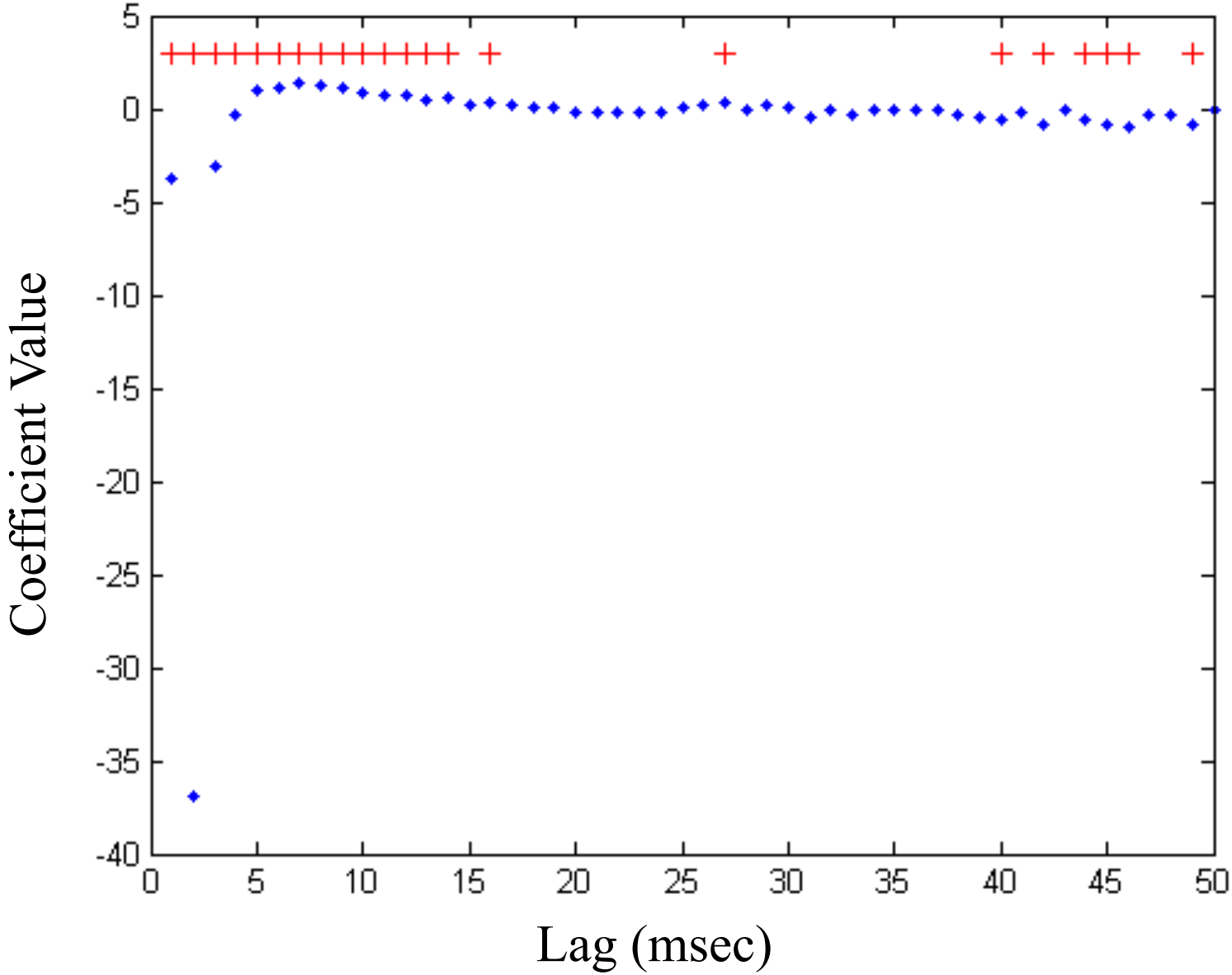
Now, let the conditional intensity be a function of past spiking activity using GLM

$$\lambda_k = \exp\left\{\alpha_0 + \sum_{i=1}^{order} \alpha_i dN_{k-i}\right\}$$

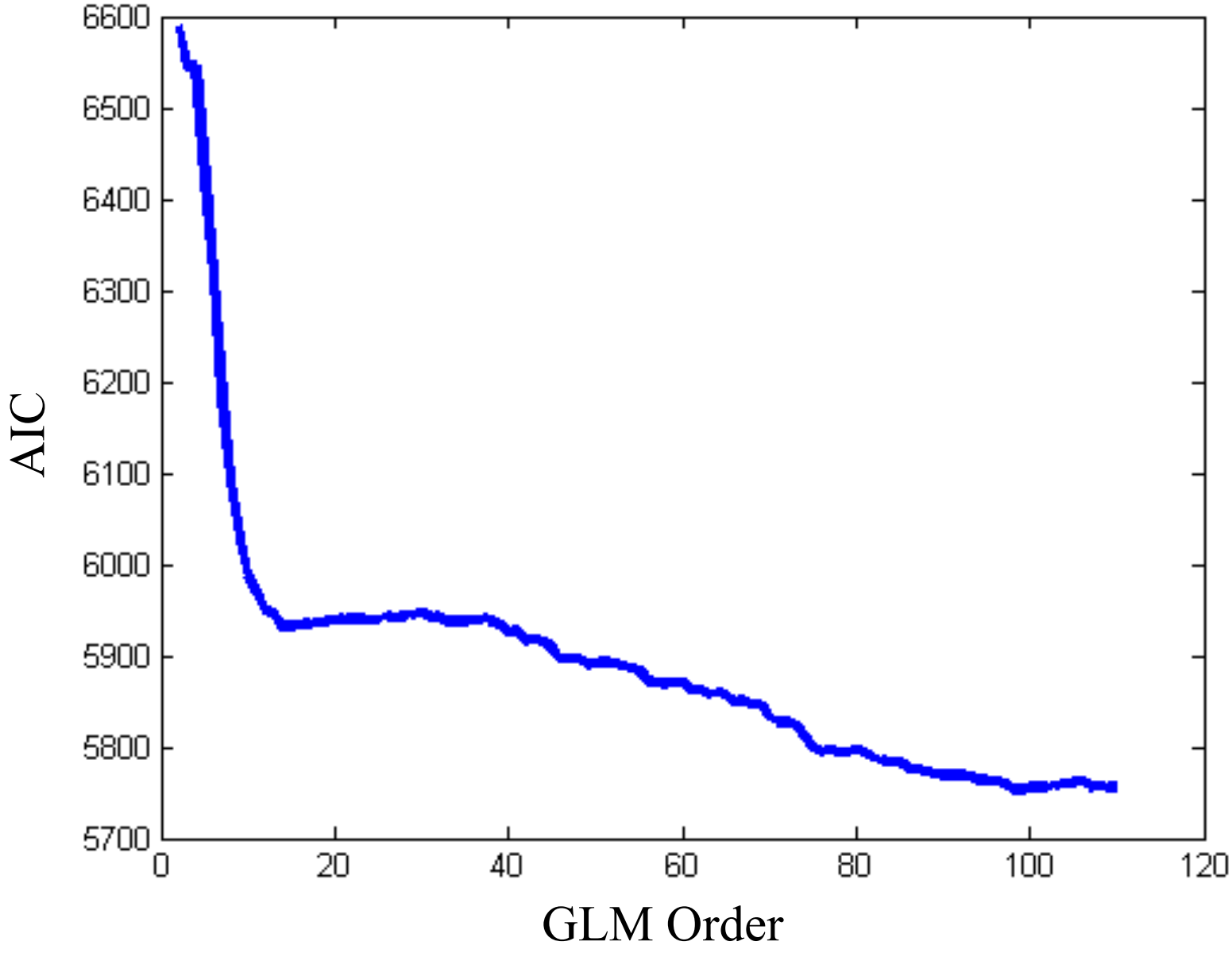
How do we pick a model order?


Partial Correlogram
Partial Correlation Coefficient
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
0
10
20
30
40
50
Model Order

GLM Coefficients

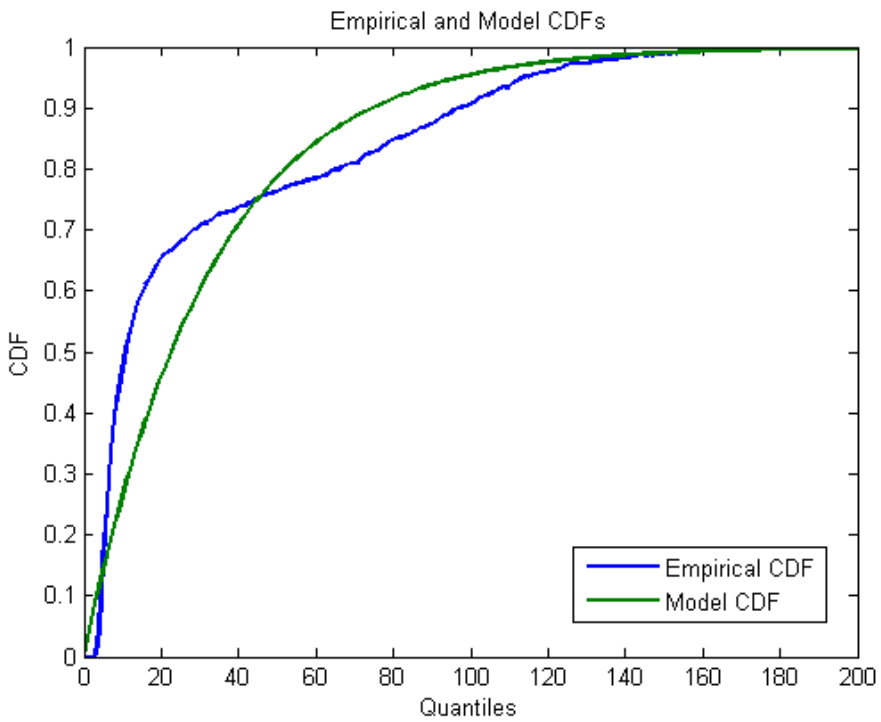


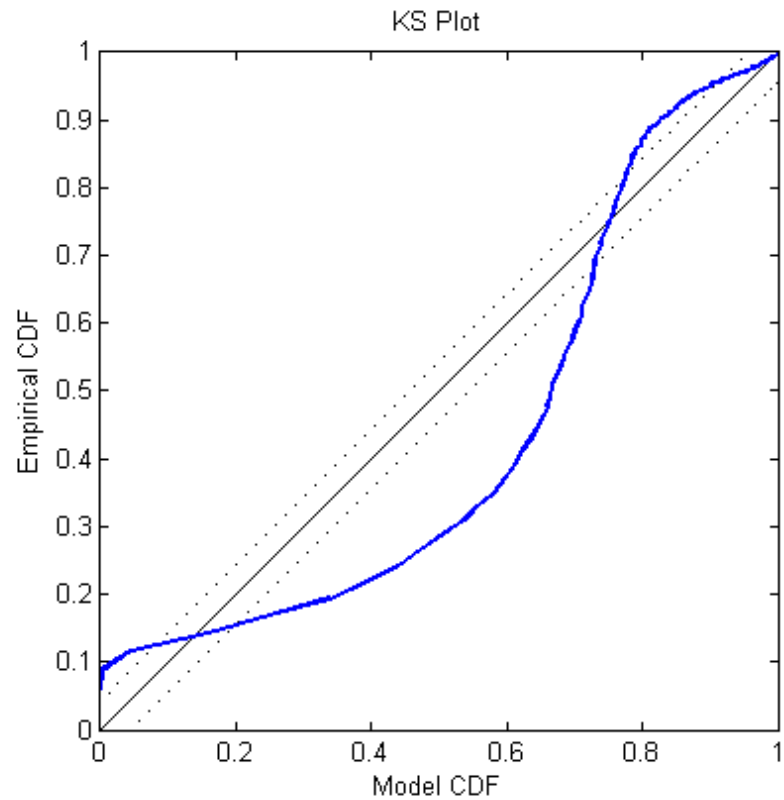
AIC Results


6600
6500
6400
6300
6200
6100
6000
5900
5800
5700
AIC
0
20
40
60
80
100
120
GLM Order

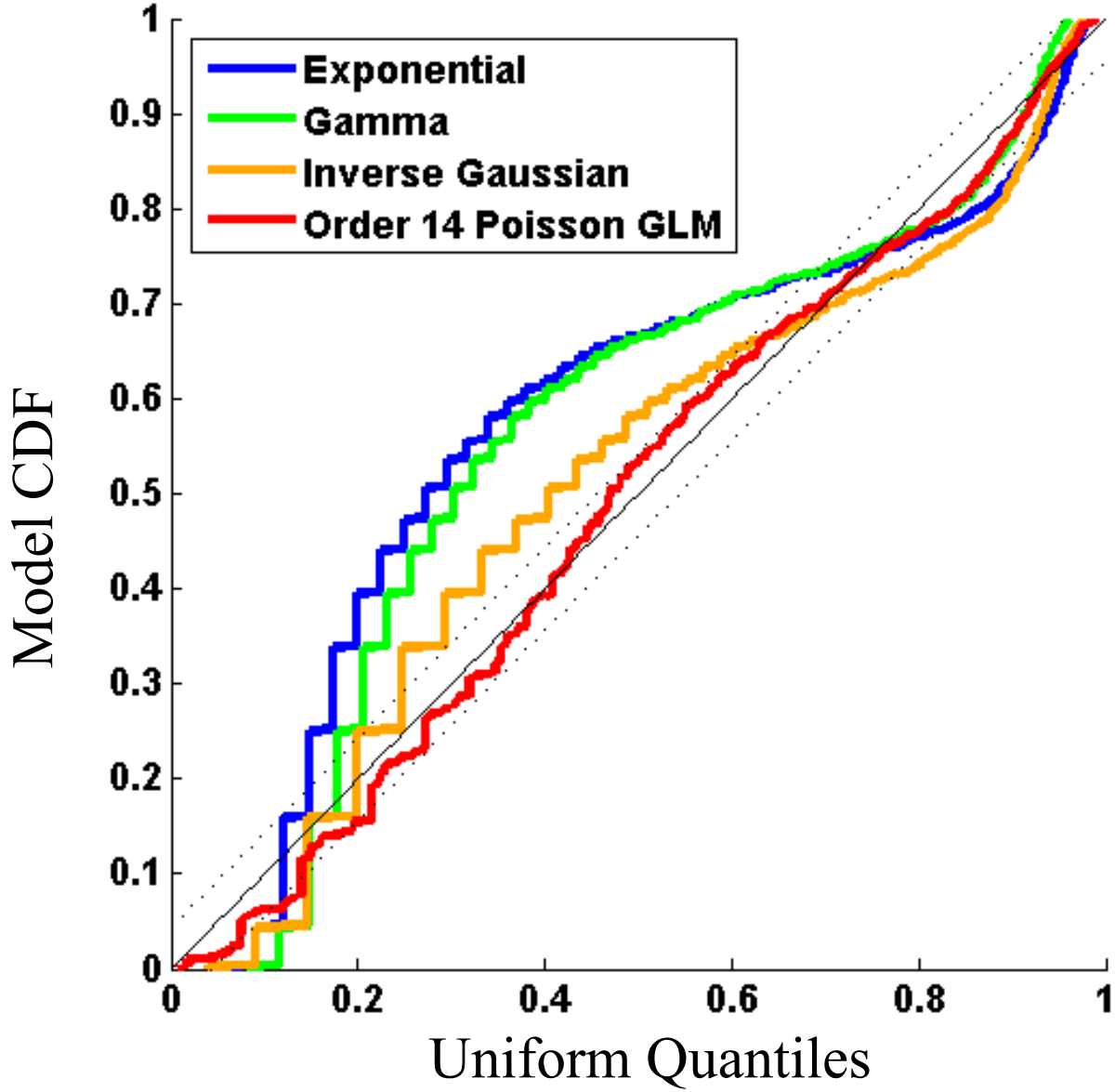
KS Plots

Graphical measure of goodness-of-fit, based on time rescaling, comparing an empirical and model cumulative distribution function. If the model is correct, then the rescaled ISIs are independent, identically distributed random variables whose KS plot should produce a 45° line [*Ogata, 1988*].

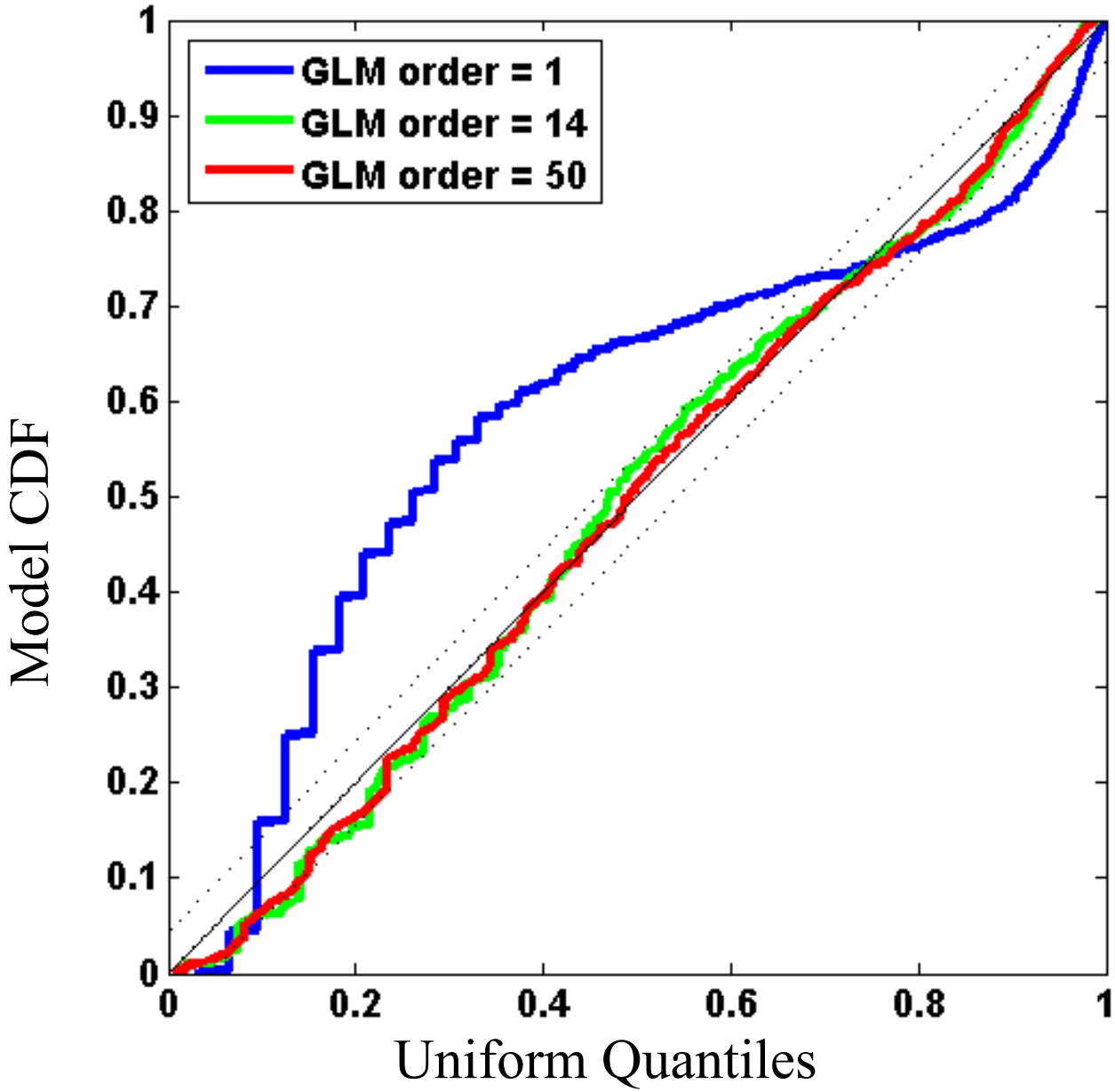


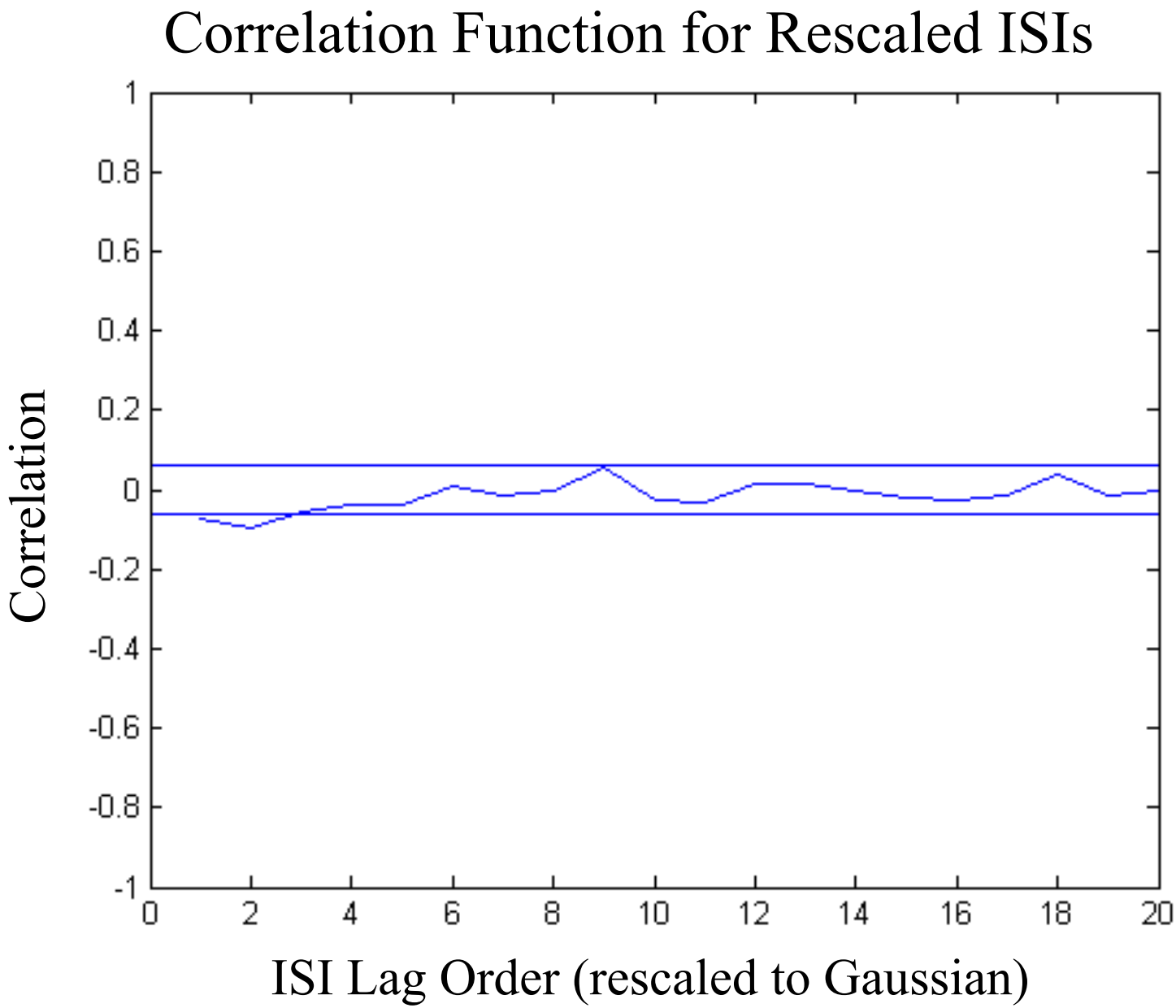


Kolmogorov-Smirnov Plots

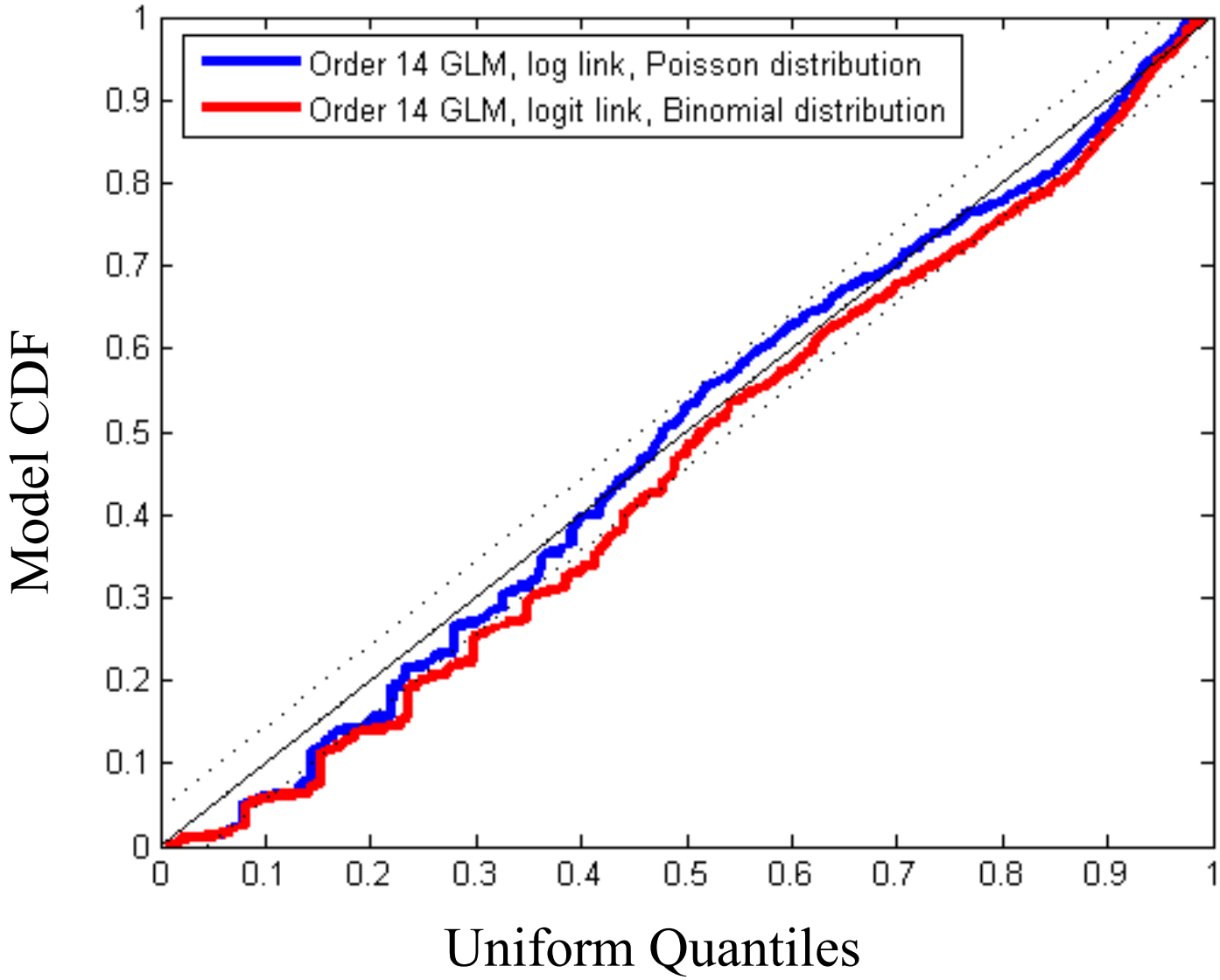


KS Plots for Different Order GLMs




Correlation Function for Rescaled ISIs
Correlation
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
2
4
6
8
10
12
14
16
18
20
ISI Lag Order (rescaled to Gaussian)

GLM Model Classes

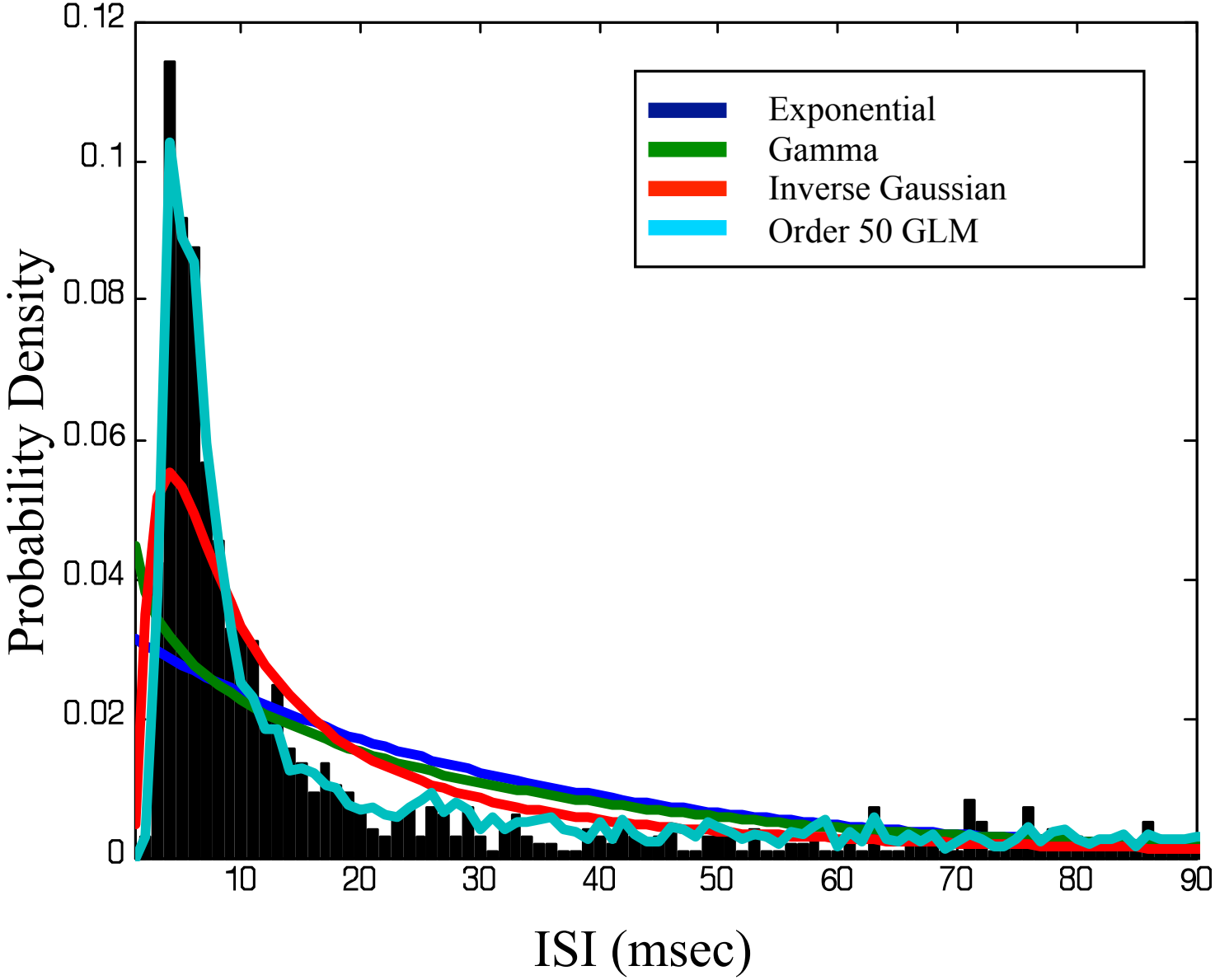


AIC and KS Statistics

	Poisson			Binomial		
Order	1	14	50	1	14	50
AIC	6589	5931	5892	8496	7792	7746
KS	0.2525	0.0657	0.0462	0.2525	0.0822	0.0533

Parametric ISI Models:	Exp	Gamma	Inv. Gauss.
KS Statistic	0.2525	0.2171	0.1063

ISI Histogram



Inferences and Conclusions

Iyengar and Liu showed that a generalized inverse Gaussian model described these data well.

The fit of history-dependent GLM model improves appreciably on the fits of the exponential, gamma and inverse Gaussian models, most notably in terms of KS plots.

Our analysis shows that the GLM model describes the essential stochastic features in the data. There is a significant history dependence in the retinal neural spiking data extending back 14 msec.

There is another effect going back approximately 100 msec.

The shorter time-scale phenomena may reflect intrinsic dynamics of the individual neuron whereas the longer time-scale effects may also include network dynamics.

Remarks

1. **Only 14 parameters are used to fit ~ 30,000 data points!**

2. **This type of strong history dependent effect is something we have seen in neurons from a number of different brain regions, animal models and experimental protocols. It was all simply described by GLM fitting.**

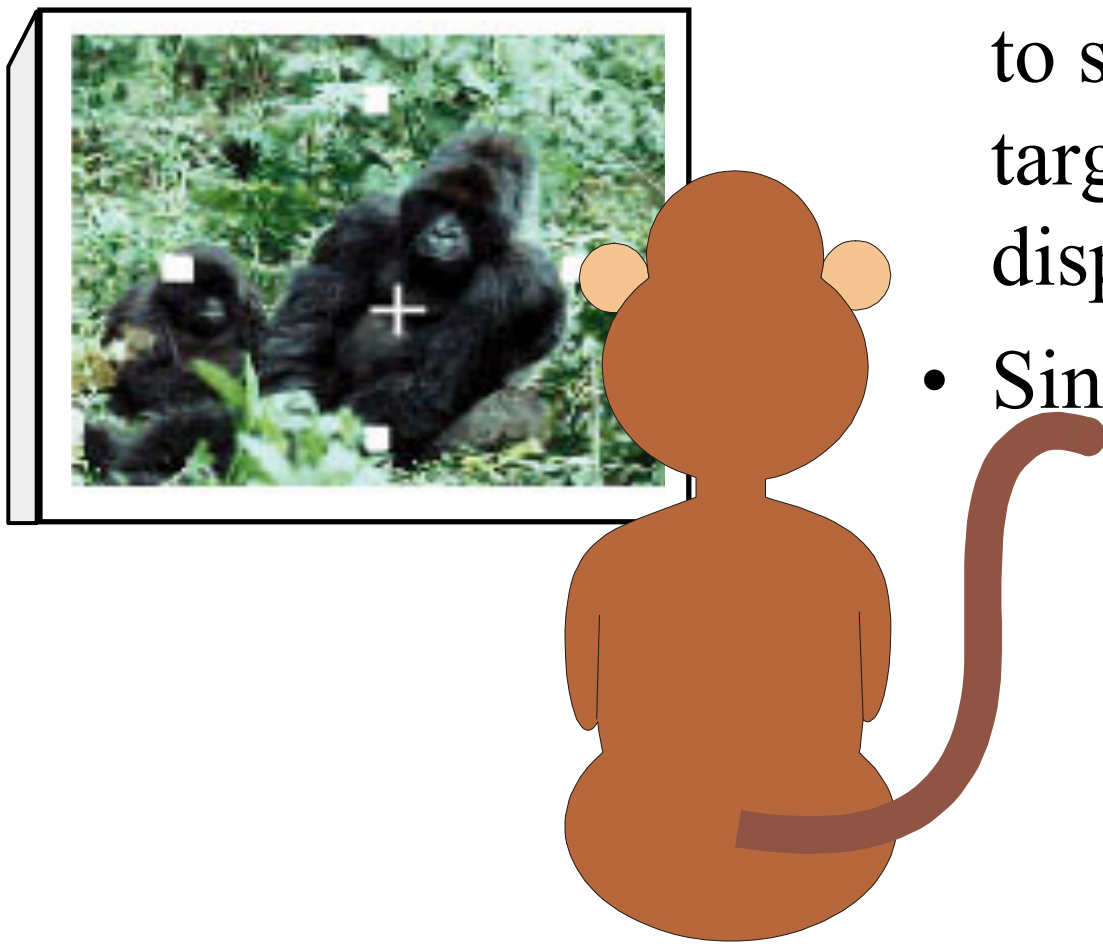
Truccolo W, Eden UT, Fellow M, Donoghue JD, Brown EN. A point process framework for relating neural spiking activity to spiking history, neural ensemble and covariate effects. *Journal of Neurophysiology*, 2005, 93:1074-1089.

Kass RE, Ventura V, Brown EN. Statistical issues in the analysis of neuronal data. *Journal of Neurophysiology*, 2005, 94: 8-25.

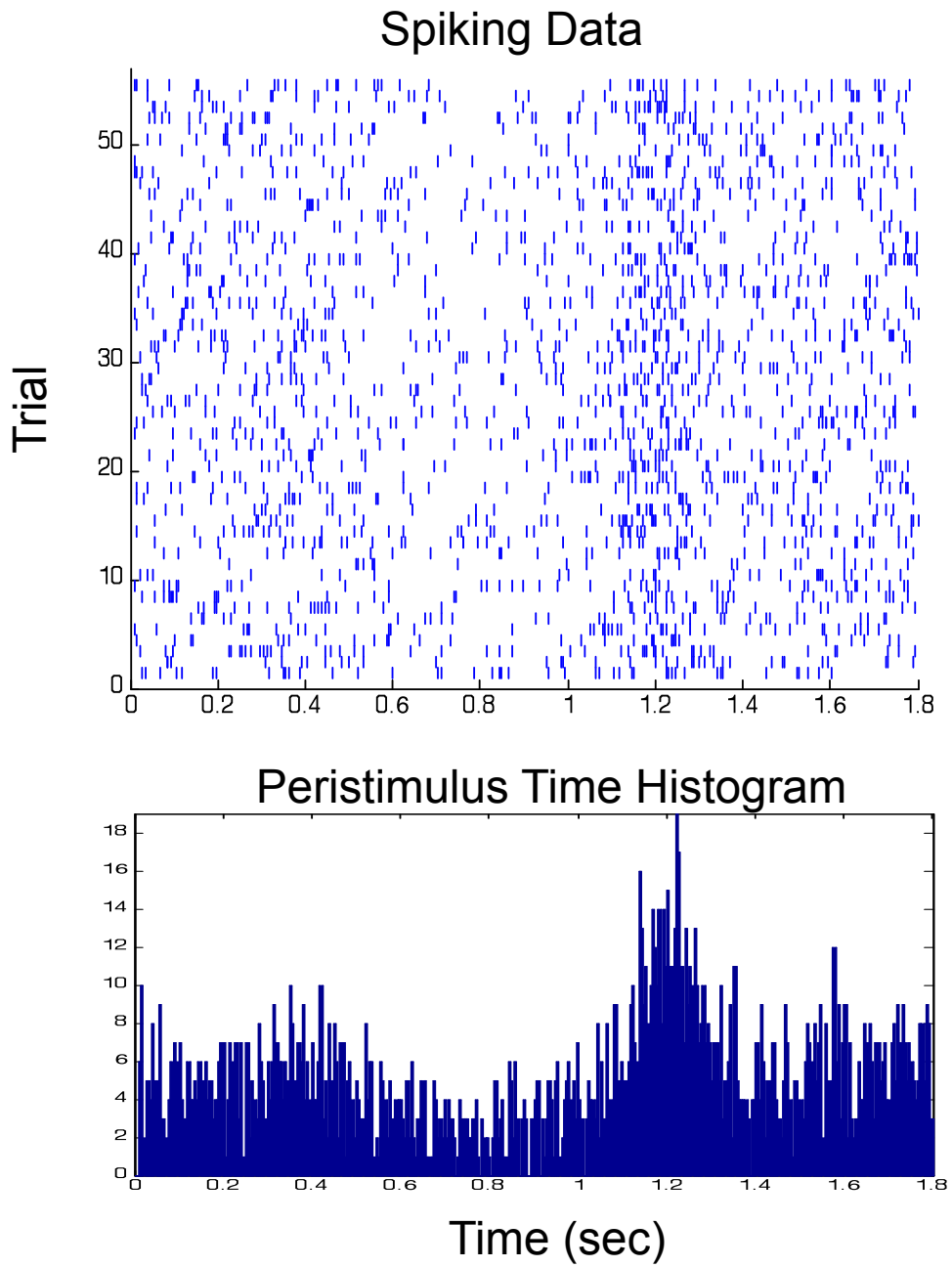
Summary

- GLM provides a computationally tractable generalization of the Gaussian linear model to non-Gaussian regression models.
- Estimation is carried out using maximum likelihood. This analysis has all the properties of maximum likelihood.
- AIC, deviance and parameter standard errors provide measures of goodness-of-fit and an inference framework analogous to regression.
- Can be applied to other exponential family models.
- Non-canonical link functions can also be used.
- GLM is a standard tool in Matlab, Minitab, R, S, SAS, Splus, and SPSS.

GLM Peristimulus Time Histogram



- Monkeys were trained to saccade to one of four targets, based on displayed images.
- Single cell recording in monkey hippocampus.

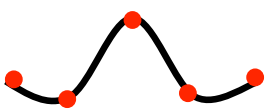

Spiking Data
Trial
0
10
20
30
40
50
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
Peristimulus Time Histogram
0
2
4
6
8
10
12
14
16
18
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
Time (sec)

Model

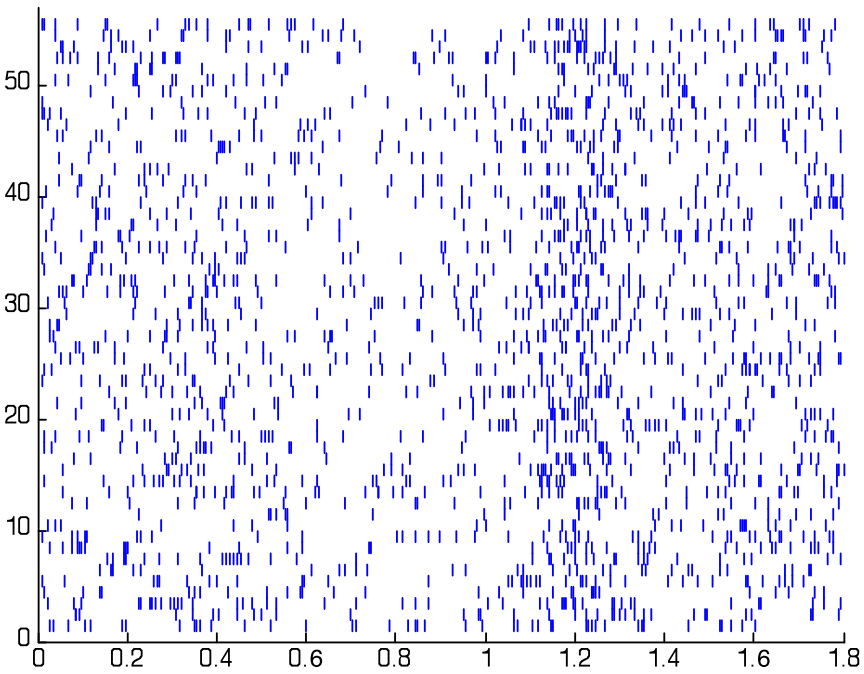
$$\lambda(t \mid H_t) = \exp\left\{ \sum_{r=1}^{R} \theta_r g_r(t) \right\}$$

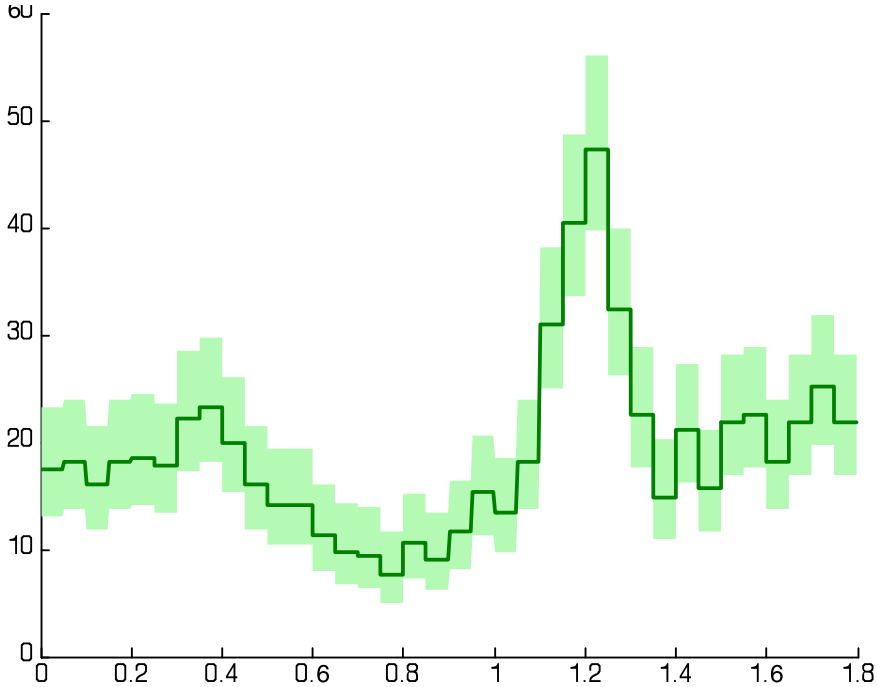
Parameter vector: $\theta = \left[\theta_1, \ldots, \theta_R\right]$

Basis functions: $g_r(t)$

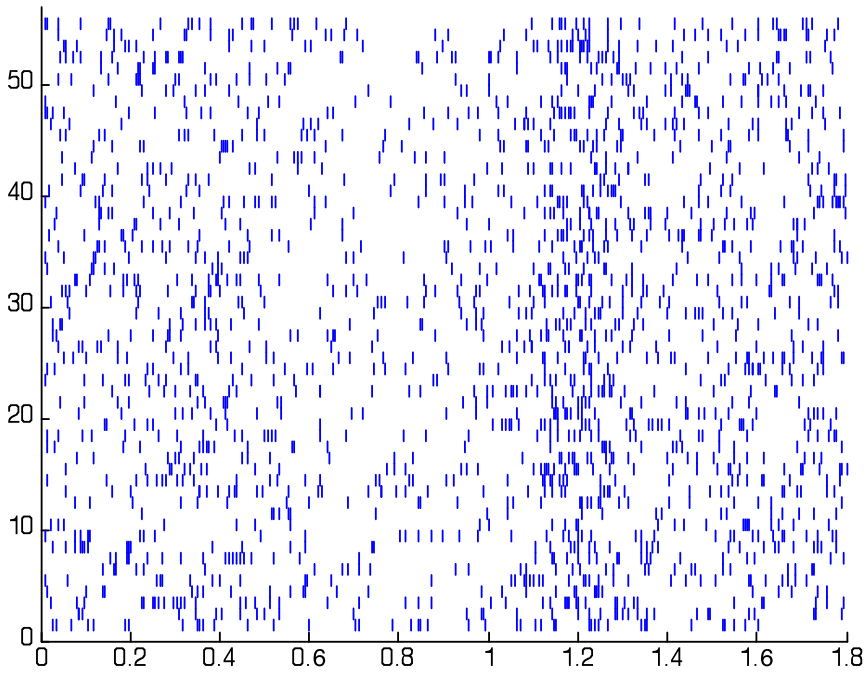
- Indicator Functions: 
- Splines: 

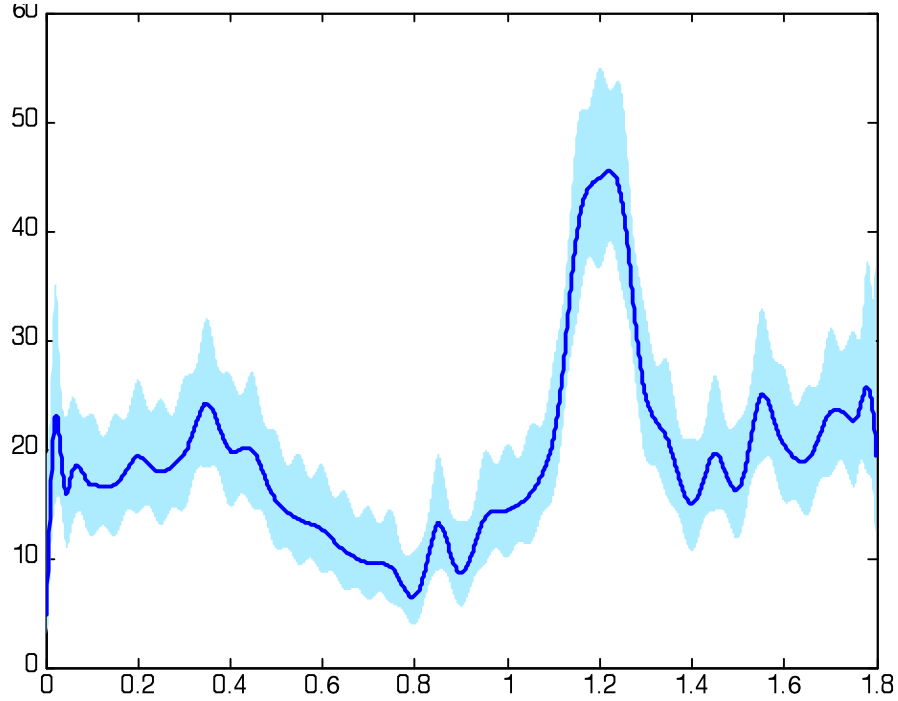
Indicator Function Basis



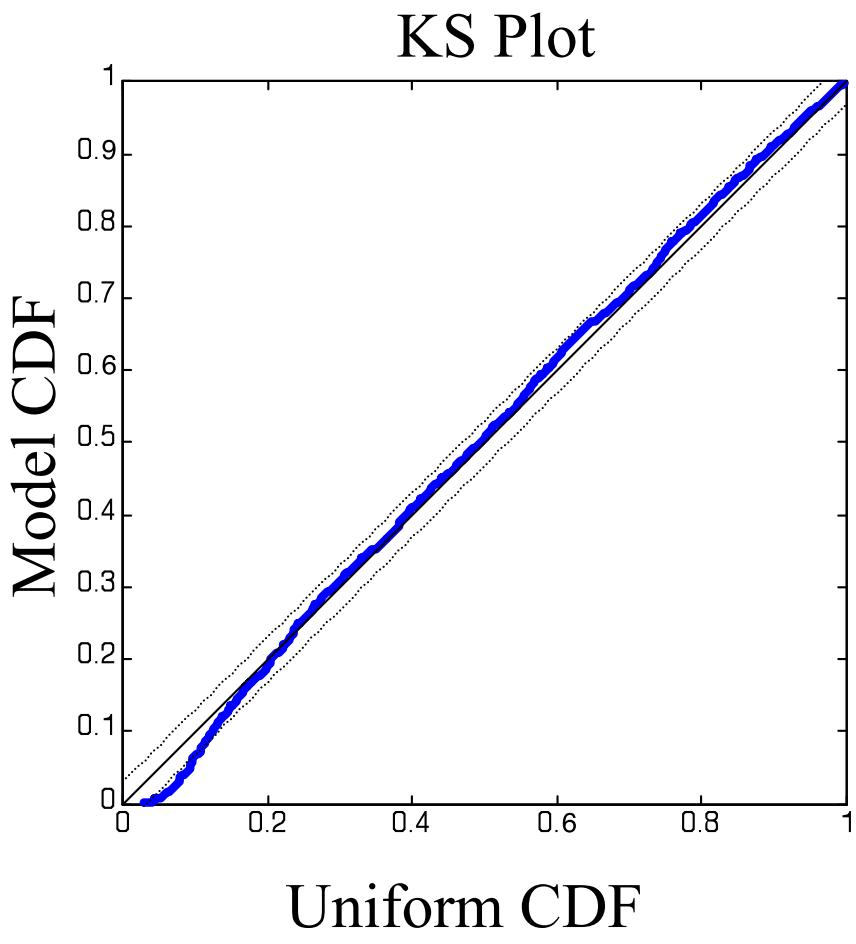


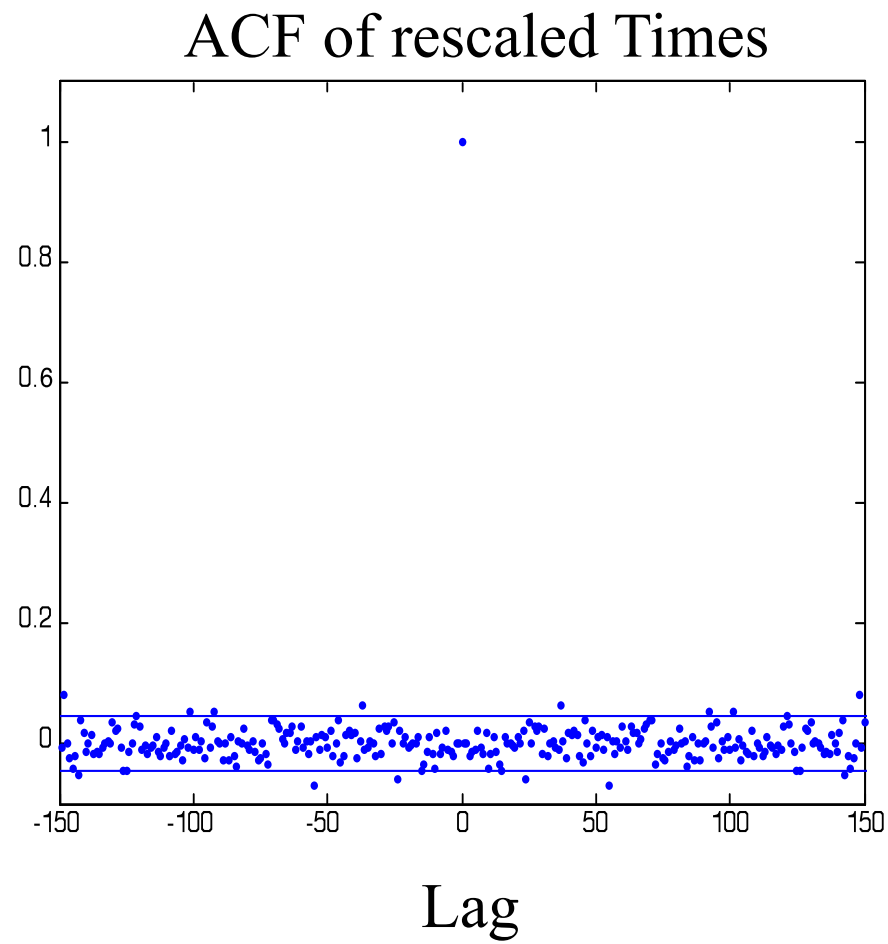
Spline Function Basis





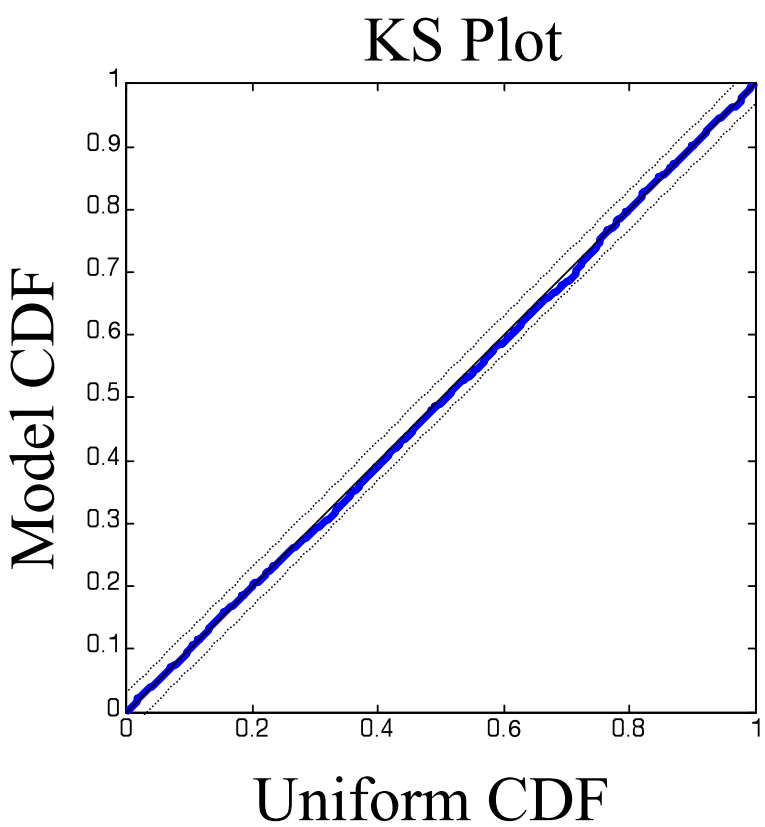
Goodness-of-Fit





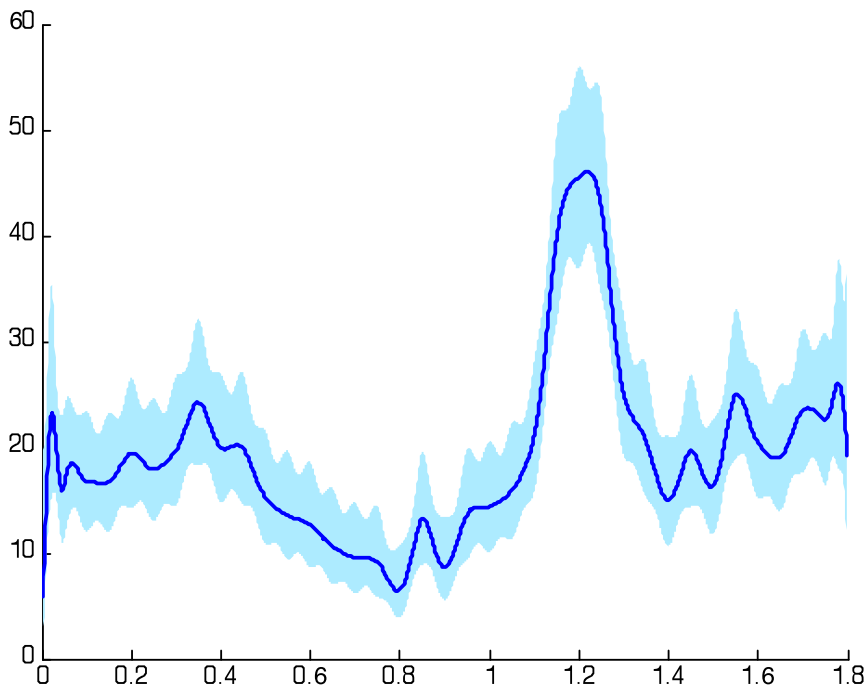
Adding History

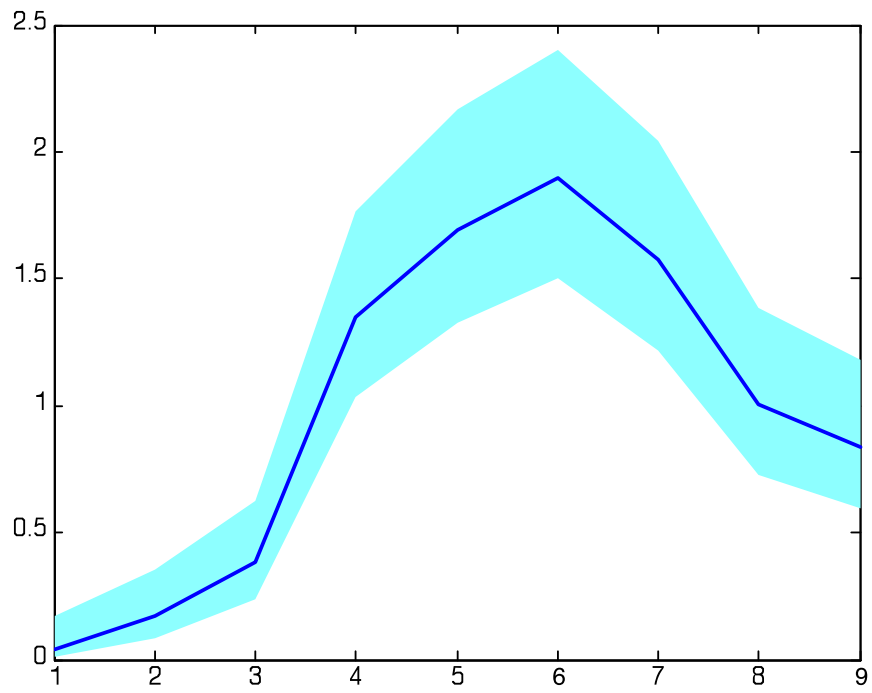
$$\lambda_k = \exp\left\{ \sum_{r=1}^{R} \theta_{k,r} g_r(t) + \sum_{i=0}^{9} \gamma_i \Delta N_{(t-i-1,t-i)} \right\}$$



Adding History

$$\lambda_k = \exp\left\{ \sum_{r=1}^{R} \theta_{k,r} g_r(t) + \sum_{i=0}^{9} \gamma_i \Delta N_{(t-i-1,t-i)} \right\}$$





GLM Neural Models

$$\log(\lambda_k) = \theta_0 + \sum_{i=1}^{I} \alpha_i f_i(\text{Extrinsic Covariates})$$

$$+ \sum_{j=1}^{J} \beta_j g_j(\text{Spiking History})$$

$$+ \sum_{k=1}^{K} \sum_{c=1}^{C} \gamma_{k,c} h_{k,c}(\text{Ensemble Activity})$$

- By selecting an appropriate set of basis functions we can capture arbitrary functional relations.
- Analysis of relative contributions of components to spiking

Truccolo W, Eden UT, Fellows MR, Donoghue JP, Brown EN. (2004) *J. Neurophys* 93:1074-1089

Conclusions

- We can construct and fit (using maximum likelihood) simple generalized linear models that capture the statistical properties of the spike train time series.
- We used the sample partial correlation function, the distribution of estimators and AIC to suggest the order of the model.
- AIC and the KS statistic are measures of goodness-of-fit between the model and the data.